A Marginal-Based Technique for Distribution Estimation

Rajasekaran Masatran MASATRAN@FREESHELL · ORG
Computer Science and Engineering, Indian Institute of Technology Madras

Abstract

A standard problem in bayesian inference is to estimate a distribution over a vector random variable, given a source of independent random instances drawn from the distribution. Frequently, the components have limited dependency between each other, and this can be exploited for estimation with fewer samples. We propose a novel technique that estimates the distribution efficiently, using two-dimensional marginals. Our technique is suited to incremental estimation. Compared to the naive bayes assumption, our technique provides better accuracy, but at higher computational cost. Experiments on datasets of different dimensionality support our claims.

1. Introduction

In bayesian inference, a standard problem is to estimate a distribution over a vector random variable in S^d, given a source of independent random instances drawn from the distribution. Here, S is a nominal attribute. Each instance is a vector of ordered components. A real-valued attribute can be converted into a nominal-valued attribute by binning. Frequently, the components have limited dependency between each other, and it is necessary to exploit this for estimation to be tractable in number of samples.

1.1. Motivation: Estimating the Distribution

Distribution estimation is necessary for bayesian classification. The problem instance to be classified is a vector $X = (x_1, x_2, \cdots x_d)$. This belongs to one of a set of classes C_k. We have the prior probabilities for C_k, possibly computed from the training data. The class-conditional distributions are learned from training data. Given the prior probabilities and the class-conditional distributions, we use bayes' theorem to compute the *posterior probabilities* $P(C_k|X)$. These are used to assign instances of test data to classes.

$$p(C_k|X) = \frac{p(C_k) \times p(X|C_k)}{p(X)} \tag{1}$$

1.2. Motivation: Two-Dimensional Marginals

A graphical explanation of how 2-D histograms capture information that is absent in 1-D histograms is in Figure 1 2-Dimensional Marginals. This is similar to Figure 10.10 *MAP solution vs. max marginals solution* from (Prince, 2012). This figure shows the joint distribution $p(x_1, x_2)$ and marginal distributions $p(x_1)$ and $p(x_2)$. The objective is to estimate the joint distribution and locate its maximum. This can be done using 1-D or 2-D marginals. Using 1-D marginals reduces the variance component of error significantly. On the other hand, using 2-D marginals is superior when the components are far from (conditionally) independent, as in this case. When 1-D histograms are used, the maximum is wrongly estimated as $(x_1, x_2) = (3, 3)$, boxed in the figure. When a 2-D histogram is used, the maximum is correctly estimated as $(x_1, x_2) = (1, 1)$, boxed in the figure.

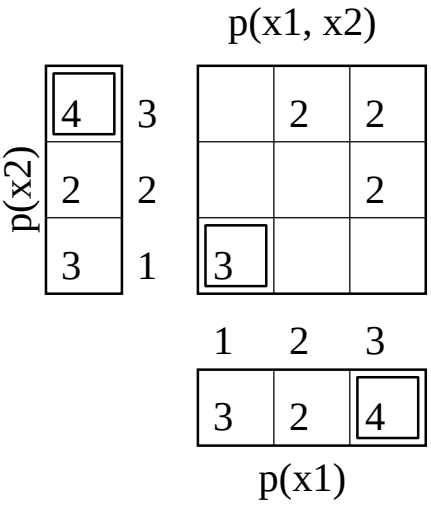

Figure 1. 2-Dimensional Marginals

2. Background

The estimation algorithm creates a representation of the distribution using training data. The querying algorithm uses this representation to estimate the probability of test data.

2.1. Prior Work

The use of one-dimensional marginal histograms in bayesian classification is discussed in subsection 6.6.3 *The Naive Bayes Classifier* of (Hastie et al., 2009).

(Pazzani, 1996) explored ways to improve naive bayes by searching for dependencies among attributes. (Domingos & Pazzani, 1997) showed that naive bayes performs well even when the independence assumption is violated, and that detecting attribute dependence is not necessarily the best way to extend the bayes classifier. (Wang et al., 2016) describes geometric density estimation (GEODE), which assumes a low-dimensional structure in the data and uses Probabilistic PCA (PPCA) for distribution estimation.

(Webb et al., 2005) describes aggregating one-dependence estimators (AODE), which weakens the attribute independence assumption by averaging all of a constrained class of classifiers. (Eban et al., 2014) takes 1-D and 2-D marginal distributions as given and chooses that joint distribution which has minimum worst-case error in classification. These two techniques cannot be directly compared with ours since ours is not restricted to classification.

2.2. Attribute Independence Assumption

The naive bayes classifier is a simple classifier that uses bayes' theorem and the attribute independence assumption. By this assumption, the features are independent within each class. $p(X|C_k) = \prod_i p(x_i|C_k)$. Each of the d components is a feature, and the features are approximately independent. The training set consists of a set of instances for each class. This set is used to estimate the class-conditional density $p(X|C_k)$ for each class C_k by computing the one-dimensional histograms. Our technique improves upon the attribute independence assumption for estimating the class-conditional density.

2.3. Framework

We use a common framework for the full-dimensional histogram and the set of one-dimensional histograms. c is the number of dimensions to be considered at a time, and varies from 1 to d. For each c, the distribution is represented by the $\binom{d}{c}$ histograms. The task is to design a technique that uses these $\binom{d}{c}$ histograms to estimate the probability at any given query point. For $c = 1$ and $c = d$, the standard techniques apply.

2.4. Bias-Variance Tradeoff

The bias-variance tradeoff is determined by the number of parameters, which in histograms is determined by the number of bins. Assuming that all histograms have m bins along each dimension, a c-dimensional histogram has m^c bins. Thus, variance increases and inductive bias decreases, with increase in c. In this paper, we design a technique for the $c = 2$ case.

When we have too many samples, the standard solution is to compute a full-dimensional histogram. When we have too few samples, the standard solution is to compute the set of one-dimensional component histograms. The attribute independence assumption is used to approximate the joint distribution as the product of these marginals. In terms of number of samples, there is a large range between the full-dimensional histogram and the set of one-dimensional marginal histograms. In this range, the full-dimensional histogram has too much error due to variance, and the set of one-dimensional marginal histograms has too much error due to inductive bias. We design a technique for this range of number of samples, that uses two-dimensional marginal histograms.

2.5. Techniques and Contributions

We analyze three techniques in this paper. While **T1** is the existing technique, **T2** and **T3** are our contributions.

T1 1-D marginal histograms are used.

T2 2-D marginal histograms are used.

T3 1-D and 2-D marginal histograms are merged.

3. Technique

The problem is to estimate the probability at the given point using the two-dimensional histograms. We do this by computing the *specific marginal* for each component, at the given point, and multiplying the specific marginals to estimate the probability at the given point.

3.1. Notation

X Vector-valued random variable, takes values in $[0, 1]^d$

d The dimensionality of the space of X

n The cardinality of the training dataset

M The training dataset, matrix of size $n \times d$, separate for each class

$q(x)$ The distribution to be learnt

$\hat{q}_j(x_j)$ The one-dimensional marginal of X, with dimension j

$\hat{q}_{j,k}(x_j, x_k)$ The two-dimensional marginal of X, with dimensions j and k

$\hat{q}(x)$ The representation of the approximation of $q(x)$

$\hat{q}_j(x_j)$ The specific marginal of $q(x)$ specific to x computed by technique T2

$s_j(x_j)$ The specific marginal of $q(x)$ specific to x computed by technique T3

3.2. Histograms

Our technique works on vectors that contain combination of nominal, ordinal and cardinal attributes. Without loss of generality, we assume that the domain of the input is identical in all dimensions. Binning is frequently required, since (1) Histograms are computed over countable sets, and (2) Error due to variance is reduced by binning.

Nominal attributes trivially fit into this framework. Hierarchical agglomeration is required if there are too many classes.

Ordinal attributes trivially fit into this framework. Ordered binning is usually required as there are too many classes.

Cardinal attributes require binning. The choice of number of bins is critical. If we use too few bins, error due to inductive bias will hurt overall performance. If we use too many bins, error due to variance will hurt overall performance.

3.3. Specific Marginal

For technique T2, the specific marginal $\hat{q}_j(x_j)$ of $q(x)$ specific to y is the geometric mean of the square root of the two-dimensional marginals that include dimension j. This differs from marginals used in naive bayes, in that this is specific to the point X for which it is computed.

$$\hat{q}_j(y_j) \leftarrow \left(\prod_k^d \sqrt{\hat{q}_{j,k}(y_j, y_k)} \right)^{\frac{1}{d}} \tag{2}$$

For technique T3, the specific marginal $s_j(y_j)$ of $q(X)$ specific to y is the weighted geometric mean of specific marginal $\hat{q}_j(x_j)$ from technique T2, and $q_j(y_j)$, with weights β and $1-\beta$ respectively. The experiments were done with $\beta = 0.50$ everywhere.

$$s_j(y_j) \leftarrow \hat{q}_j(y_j)^{\beta} q_j(y_j)^{1-\beta} \tag{3}$$

3.4. Estimation

Given a set of samples M from the distribution with semi-dependent marginals $q(X)$, compute $\hat{q}(X)$, a representation of $q(X)$.

Algorithm 1: Estimation ($O(n\,d^2)$)

Input: M $(n \times d)$
Output: $\hat{q}_{i,j}$
for $i \in \{1 \cdots n\}$ **do** $\underline{O(n\,d^2)}$
  for $j \in \{1 \cdots d\}$ **do** $\underline{O(d^2)}$
    for $k \in \{1 \cdots d\} \setminus \{j\}$ **do** $\underline{O(d)}$
      increment $\hat{q}_{j,k}(M_{i,j}, M_{i,k})$
      $\underline{\hat{q}_{j,k} \text{ is the marginal with dimensions } j \text{ and } k}$
    end
  end
end

Algorithm 1: We compute the two-dimensional marginals from the given samples. This representation is used by the querying algorithm to estimate the probability density at the query point y.

3.5. Shrinkage

We take the weighted mean of the representation histogram and the uniform distribution, with weights $(1-\alpha)$ and α respectively, for smoothing. This is standard procedure to ensure that error due to variance in zero-valued bins does not skew the results. The experiments were done with $\alpha = 0.05$ everywhere.

3.6. Querying

Given the representation $\hat{q}(X)$ of a distribution $q(X)$, and y, estimate $q(y)$.

Algorithm 2: Querying ($O(d^2)$)

Input: $\hat{q}_{i,j}, y$
Output: $\hat{q}(y)$
for $j \in \{1 \cdots d\}$ **do** $\underline{O(d^2)}$
  $\hat{q}_j(y_j) \leftarrow \left(\prod_k^d \sqrt{\hat{q}_{j,k}(y_j, y_k)} \right)^{\frac{1}{d}}$ $\underline{O(d)}$
  $s_j(y_j) \leftarrow \hat{q}_j(y_j)^{\beta} q_j(y_j)^{1-\beta}$
  $\hat{q}(y) \leftarrow \prod_j^d s_j(y_j)$ $\underline{O(d)}$
end

Algorithm 2: We compute all the specific marginals of y. We multiply them together to estimate the probability, just as we would if we were approximating the joint distribution with the product of its one-dimensional marginals.

3.7. Analysis

Consider the case where all components are independent of each other. In this trivial case, the two-dimensional histogram technique reduces to the product of one-dimensional marginals. Therefore, our assumption is weaker than

the attribute independence assumption.

$$
\begin{aligned}
s_j(y_j) &= \hat{q}_j(y_j)^{\beta} q_j(y_j)^{1-\beta} \\
&= \left(\prod_k^d \sqrt{\hat{q}_{j,k}(y_j, y_k)} \right)^{\frac{\beta}{d}} q_j(y_j)^{1-\beta} \\
&= \left(\prod_k^d \sqrt{\hat{q}_j(y_j)\hat{q}_k(y_k)} \right)^{\frac{\beta}{d}} q_j(y_j)^{1-\beta} \\
&= \left(\prod_k^d \sqrt{\hat{q}_k(y_k)} \right)^{\frac{\beta}{d}} q_j(y_j)^{1-\frac{\beta}{2}}
\end{aligned}
$$

$$
\begin{aligned}
s(y) &= \prod_j^d s_j(y_j) \\
&= \prod_j^d \left(\left(\prod_k^d \sqrt{\hat{q}_k(y_k)} \right)^{\frac{\beta}{d}} q_j(y_j)^{1-\frac{\beta}{2}} \right) \\
&= \left(\prod_k^d \hat{q}_k(y_k) \right)^{\frac{\beta}{2}} \left(\prod_j^d \hat{q}_j(y_j) \right)^{1-\frac{\beta}{2}} \\
&= \prod_j^d \hat{q}_j(y_j)
\end{aligned}
$$

3.8. Time Complexity

Algorithm	1-D	2-D	Merged
Estimation	$O(n\,d)$	$O(n\,d^2)$	$O(n\,d^2)$
Querying	$O(d)$	$O(d^2)$	$O(d^2)$

Table 1. Time Complexity

4. Experiments

ISOLET (Isolated Letter Speech Recognition) is a dataset of features of 150 speakers saying the 26 letters of the English alphabet twice (Fanty & Cole, 1991). 120 speakers were recorded for training, and 30 for testing.

The algorithms were coded in Octave and run on Linux. The representation generated during estimation was saved to disk in the estimation algorithm, and reloaded in the querying algorithm.

The experiments were run on a computer with an Intel 5200U processor and 12 gigabytes of memory. During the experiments, processor utilization was the bottleneck, and memory utilization was minimal.

4.1. Data

Dimensions 617, all real, normalized to $[-1, +1]$

Classes 26

Data 300 instances per class, 3 missing instances

Training Data 240 instances per class

Test Data 60 instances per class

4.2. Preprocessing

To compute the histogram, we need the domain of the input to be a countable set. So, we bin the input, using 16-bin one-dimensional histograms and 4×4–bin two-dimensional histograms. Since the number of bins is equal in both, the bias-variance tradeoff in both systems is roughly equal. Thus, our input is transformed from $\mathbb{R}^d$ to S^d, and S has cardinality 4.

4.3. Evaluation

We evaluated techniques T1, T2, and T3 for different numbers of dimensions. For each dimension count, the dimensions were selected at random, and the evaluation was performed on this dimension set for each of the three techniques: 1-D histograms, 2-D histograms, and the merged technique.

4.4. Caveats

1. In estimation, for each dimension count, the set of dimensions was selected only once.
2. In querying, only the first 10% of test data for each class was used.
3. The shrinkage coefficient α and the marginal weight β must be selected by cross-validation.

4.5. Results

Dimensions	1-D	2-D	Merged
4	0.145	0.143	0.150
16	0.547	0.519	0.588
64	0.826	0.790	0.825
256	0.916	0.872	0.911

Table 2. Accuracy

Our technique works as expected, providing better accuracy, at higher computational cost. For up to 2^2 dimensions, the full-dimensional histogram will provide the best results on this dataset. For more than 2^6 dimensions, naive bayes will provide the best results on this dataset. From 2^3 to 2^5 dimensions, our technique provides the best results among these three techniques on this dataset.

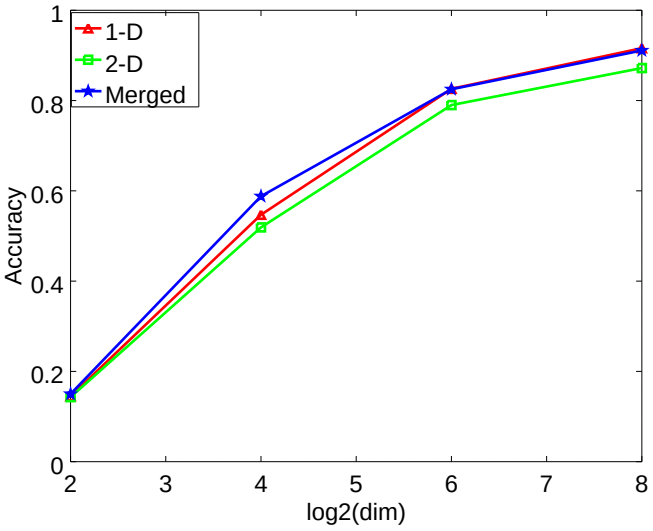

Figure 2. Accuracy

5. Conclusion

We have proposed a technique for estimating a distribution over a vector random variable that uses samples from the distribution efficiently. Experiments support our claims for datasets of different dimensionality. Compared to the attribute independence assumption, our technique provides better accuracy, but at higher computational cost.

5.1. Future Work

1. Prove convergence in the attribute independence case

2. Analyze the variance of the three techniques

Acknowledgments Chiranjib Bhattacharyya, Pavan Mallapragada, Prasanna Parthasarathy.

References

Domingos, Pedro and Pazzani, Michael J. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29(2):103–130, 1997.

Eban, Elad, Mezuman, Elad, and Globerson, Amir. Discrete Chebyshev Classifiers. In *ICML 2014: Proceedings of the Thirty-First International Conference on Machine Learning*, pp. 1233–1241, 2014.

Fanty, Mark and Cole, Ronald. Spoken Letter Recognition. In *NIPS 1991: Proceedings of the Third International Conference on Neural Information Processing Systems*, pp. 220–226, 1991. URL https://archive.ics.uci.edu/ml/datasets/ISOLET.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, second edition, 2009. ISBN 978-0-387-84857-0. URL http://statweb.stanford.edu/~tibs/ElemStatLearn/.

Pazzani, Michael J. Searching for Dependencies in Bayesian Classifiers. In *AISTATS 1995: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pp. 239–248, 1996.

Prince, Simon J. D. *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012. ISBN 1107011795, 9781107011793.

Wang, Ye, Canale, Antonio, and Dunson, David B. Scalable geometric density estimation. In *AISTATS 2016: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics*, pp. 857–865, 2016.

Webb, Geoffrey I., Boughton, Janice R., and Wang, Zhihai. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*, 58(1):5–24, 2005.