A Marginal-Based Technique for Distribution Estimation

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

Abstract

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

1. Introduction

In statistics, 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 aspect 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.

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

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 principal component analysis (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, and computes the joint distribution that 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. $p(X|C_k) = \prod_j p(x_j|C_k)$. 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.

3. Technique

The problem is to estimate the probability at the given point using the one-dimensional histograms.

$$p(X|C_k) = \prod_j p(x_j|C_k)$$

$$\log(p(X|C_k)) = \sum_j 1 \log(p(x_j|C_k))$$

The above is naive bayes (NB), the orignial technique. Our technique is modified bayes (MB). Here, instead of directly adding up the log probabilities, we assign a coefficient to each log probability, and do regression on the training data to estimate the coefficients. Originally, the coefficients can be taken to be 1 each. Our coeffients do not need to add up to d.

$$\log(p(X|C_k)) = \sum_j \alpha_j \log(p(x_j|C_k))$$

3.1. Argument

Modified bayes assumes that the assumptions of naive bayes are problematic. We start with replicating naive bayes, and argue that it can handle irrelevant attributes, and one type of non-independent attributes.

Generalization of NB In case the assumptions of naive bayes are valid, regression coeffients $\alpha_j = 1$ replicate the naive bayes formula.

Irrelevant Attributes In case irrelevant attributes are added, then the corresponding regression coeffients can be $\alpha_j = 0$.

Dependenct Attributes In case an attribute is replaced by three noisy duplicates, then the corresponding regression coeffients for the three can be $\sum_j \alpha_j = 1$.

3.2. Contributions

We analyze two techniques in this paper. While **NB** is the existing technique, **MB** is our contribution.

NB 1-D marginal histograms are used, via naive bayes.

MB 1-D marginal histograms are used, via modified bayes.

3.3. 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

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

3.4. Histograms

Our technique works on vectors that contain a 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.

3.5. 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 ($O(n\,d)$)

```
Input: $M$ ($n \times d$)
Output: $\hat{q}_j$
for $i \in \{1 \cdots n\}$ do                                $\underline{O(n\,d)}$
    for $j \in \{1 \cdots d\}$ do                             $\underline{O(d)}$
        increment $\hat{q}_j(M_{i,j})$
            $\underline{\hat{q}_j \text{ is the marginal with dimension } j}$
    end
end
```

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

3.6. Shrinkage

We take the weighted mean of the representation histogram and the uniform distribution, with weights $(1 - \gamma)$ and γ 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 are done with $\gamma = 0.05$ everywhere.

3.7. Regression

$$\log(p(X|C_k)) = \sum_j \alpha_j \log(p(x_j|C_k))$$

$$\sum_j \log(p(x_j|C_k))\alpha_j = \log(p(X|C_k))$$

This is of the form

$$MA = Y$$

The solution is:

$$A = M^+ Y$$

where M^+ is the pseudo-inverse of M.

$$M^+ = (M'M)^{-1}M'$$

Using associativity might give improved accuracy in the computation of M^+Y.

3.8. Querying

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

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

```
Input: $\hat{q}_j, y$
Output: $\hat{q}(y)$
for $j \in \{1 \cdots d\})$ do                                  $\underline{O(d^2)}$
    $\log(\hat{q}(y)) \leftarrow \sum_j \alpha_j \log(\hat{q}_j(y_j))$          $\underline{O(d)}$
end
```

Algorithm 2: We estimate the probability of the joint distribution as the minimum of the corresponding one-dimensional marginals.

3.9. Time Complexity

Algorithm	Naive Bayes	Ours
Estimation	$O(n\,d)$	$O(n\,d)$
Regression		$O(n^2\,d)$
Querying	$O(d)$	$O(d)$

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.

Our 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. Thus, our input is transformed from $\mathbb{R}^d$ to S^d, and S has cardinality 16.

4.3. Evaluation

We evaluated techniques NB and MB 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 both naive bayes and modified bayes.

4.4. Caveats

1. In estimation, for each dimension count, the set of dimensions was selected only once.
2. The shrinkage coefficient γ must be selected by cross-validation.

4.5. Results

(to be written)

5. Conclusion

We have proposed a technique for estimating a distribution over a vector random variable that surpasses naive bayes, without increase in computational complexity. Experiments support our claims for datasets of different dimensionality.

5.1. Future Work

1. Analyze the statistical properties of the technique.

Acknowledgments Balaraman Ravindran.

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.

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.