Generic Text Categorization using Naïve Bayes

Manas Ram Bapatla
University of Texas at Arlington, Texas
manasram.bapatla@mavs.uta.edu

Rajiv Ravishankar
University of Texas at Arlington, Texas
rajiv.ravishankar@mavs.uta.edu

Abstract

Naïve Bayes is a simple Bayesian classifier has been found to work very well with text categorization. It is a probabilistic approach which makes strong assumptions about how the data is generated. It assumes that all attributes of the examples are independent of each other given the context of the class. While this assumption is clearly false in most real-world tasks, Naive Bayes often performs classification very well. This paradox is explained by the fact that classification estimation is only a function of the sign (in binary cases) of the function estimation. The Naïve Bayes classifier is usually implemented with Gaussian distribution function. However the accuracy can be further enhanced by implementing a mixture of Gaussians and histograms. Among these implementations histograms perform significantly better when the dimension of the data set is small.

Introduction

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable 'y' and a dependent feature vector x_1 through x_n, Bayes' theorem states the following relationship.

$$P(y|x1,\dots.,xn) = \frac{P(y)P(x1,\dots xn|y)}{P(x1,\dots.xn)}$$

Using the naïve independence assumption that

$$P(x_i|y,x_1,\dots.x_{i-1},x_{i+1},\dots.x_n) = P(x_i|y)$$

For all i, this relationship is simplified to

$$P(y|x1,\dots.,xn) = \frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x1,\dots.xn)}$$

Since $P(x_1,\dots,x_n)$ is constant given the input, we can use

$$\hat{y} = argmax\ P(y)\prod_{i=1}^{n}P(x_i|y)$$

Naive Bayes learners and classifiers can be extremely fast compared to more sophisticated methods. The decoupling of the class conditional feature distributions means that each distribution can be independently estimated as a one dimensional distribution. This in turn helps to alleviate problems stemming from the curse of dimensionality.

Implementation of Naïve Bayes

In this paper Naïve Bayes is implemented using three different approaches. Gaussian or normal distribution, Mixture of Gaussians and histograms are the approaches discussed and measured. For the experiment datasets of different dimensions and sizes are used, to make it more rigorous datasets which do not have tightly coupled classes are also used.

Implementation using Normal/Gaussian distribution

When dealing with continuous data, a typical assumption is that the continuous values associated with each class are distributed according to a Gaussian distribution. For example, suppose the training data contains a continuous attribute, we first segment the data by the class, and then compute the mean and variance of that continuous attribute in each class.

The Naïve Bayes classifier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class.

The result for the datasets is as follows:

Training set	No. of records	No. of training objects	No. of test objects	No. of Attributes	No. of classes	accuracy (%)
Pima-indians-diabetes	768	514	254	8	2	76.37
Yeast	1484	1000	484	8	10	17.56
satellite	6435	4435	2000	36	6	52.25
Pen-digits	10922	7494	3498	16	10	20.06

Implementation using mixture of Gaussians

Gaussian mixture is a probabilistic model for representing the presence of subpopulations within an overall population. It does not require that an identified individual observation belongs to a certain datasets sub population. A Bayesian Gaussian mixture model is commonly extended to fit a vector of unknown parameters.

P(x | class) is modeled as a mixture of Gaussians separately for each dimension of the data. The number of Gaussians for each mixture is set to 5 in our experiments.

Suppose that you are building a mixture of N Gaussians for the i-th dimension of the data. Let S be the smallest and L be the largest value in the i-th dimension among all training data. Let G = (L-S)/N. Then, you should initialize all standard deviations of the mixture to 1, and you should initialize the means as follows:

- For the first Gaussian, the initial mean should be S + G/2.
- For the second Gaussian, the initial mean should be S + G + G/2.
- For the third Gaussian, the initial mean should be S + 2G + G/2.
- ...
- For the N-th Gaussian, the initial mean should be S + (N-1)G + G/2.

The result for the datasets is as follows:

Training set	No. of records	No. of training objects	No. of test objects	No. of Attributes	No. of classes	accuracy (%)
Pima-indians-diabetes	768	514	254	8	2	77.68
Yeast	1484	1000	484	8	10	18.66
satellite	6435	4435	2000	36	6	53.23
Pen-digits	10922	7494	3498	16	10	21.69

Implementation using Histograms

To construct a histogram, the first step is to "bin" the range of values—that is, divide the entire range of values into a series of intervals—and then count how many values fall into each interval.

We then used model P(x | class) as a histogram separately for each dimension of the data (The number of bins for each histogram is 5 in our case).

Suppose that you are building a histogram of N bins for the j-th dimension of the data. Let S be the smallest and L be the largest value in the j-th dimension among all training data. Let G = (L-S)/N. Then, your bins should have the following ranges:

- Bin 0, from -infinity to S+G.
- Bin 1, from S+G to S+2G.
- Bin 2, from S+2G to S+3G.
- ...
- Bin N-1 from S+(N-1)G to +infinity.

The result for the datasets is as follows:

Training set	No. of records	No. of training objects	No. of test objects	No. of Attributes	No. of classes	accuracy (%)
Pima-indians-diabetes	768	514	254	8	2	87.35
Yeast	1484	1000	484	8	10	47.10
satellite	6435	4435	2000	36	6	53.43
Pen-digits	10922	7494	3498	16	10	46.39

Experimental Results

This section provides empirical evidence that Naïve Bayes with histograms is better than Mixture of Gaussians, which is in turn better than Gaussian distribution. The results are based on 4 datasets.

Datasets description and results

Pima Indians diabetes dataset s comprised of 768 observations of medical details for Pima indians patents. The records describe instantaneous measurements taken from the patient such as their age, the number of times pregnant and blood workup. All patients are women aged 21 or older. All attributes are numeric, and their units vary from attribute to attribute. Each record has a class value that indicates whether the patient suffered an onset of diabetes within 5 years of when the measurements were taken (1) or not (0). The dataset has 514 training objects and 254 test objects.

Yeast dataset predicts the localization of a protein. The dataset has 1000 training objects and 484 test objects. There are 8 attributes and 10 classes.

The satellite database consists of the multi-spectral values of pixels in 3x3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values. In the sample database, the class of a pixel is coded as a number.

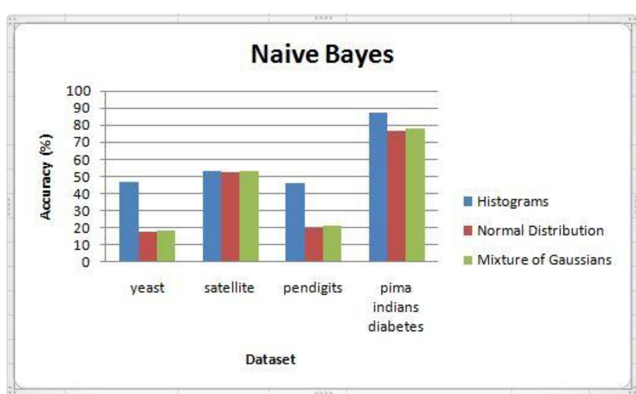
The Landsat satellite data is one of the many sources of information available for a scene. The interpretation of a scene by integrating spatial data of diverse types and resolutions including multispectral and radar data, maps indicating topography, land use etc. is expected to assume significant importance with the onset of an era characterized by integrative approaches to remote sensing (for example, NASA's Earth Observing System commencing this decade). Existing statistical methods are ill-equipped for handling such diverse data types. Note that this is not true for Landsat MSS data considered in isolation (as in this sample database). This data satisfies the important requirements of being numerical and at a single resolution, and standard maximum-likelihood classification performs very well. Consequently, for this data, it should be interesting to compare the performance of other methods against the statistical approach.

One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340 x 3380 such pixels.

The database is a (tiny) sub-area of a scene, consisting of 82 x 100 pixels. Each line of data corresponds to a 3x3 square neighborhood of pixels completely contained within the 82x100 sub-area. Each line contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighborhood and a number indicating the classification label of the central pixel. The dataset has 4435 training objects and 2000 test objects. There are 36 attributes and 6 classes.

Pen based recognition of handwritten digits dataset has 7494 training objects and 3498 test objects. There are 16 attributes and 10 classes.

Results of applying naïve bayes on these datasets are shown by the following graph.



Conclusion

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem. Among the three implementations of Naïve Bayes histogram performs significantly better when the dimensions of the dataset is small. However it does offer a small improvement over Gaussian distribution even when the dataset has fairly large dimensions. Mixture of Gaussians also performs slightly better than the normal distribution.

Future work, the implementation may be further enhanced by using feature selection and TF-IDF.

References

P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classi_er under zero-one loss. Machine Learning, 29:103{130, 1997.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classi_ers. Machine Learning, 29:131{163, 1997.

T. Kalt and W. B. Croft. A new probabilistic model of text classi_cation and retrieval. Technical Report IR-78, University of Massachusetts Center for Intelligent Information Retrieval, 1996. http://ciir.cs.umass.edu/publications/index.shtml.

Daphne Koller and Mehran Sahami. Hierarchically classifying documents using very few words. In Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

Pat Langley, Wayne Iba, and Kevin Thompson. An analysis of Bayesian classi_ers. In AAAI-92, 1992.

Leah S. Larkey and W. Bruce Croft. Combining classi_ers in text categorization. In SIGIR-96, 1996.

D. Lewis and W. Gale. A sequential algorithm for training text classi_ers. In SIGIR-94, 1994.

David D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In SIGIR-92, 1992.

David Lewis. Naive (bayes) at forty: The independence asssumption in information retrieval. In ECML'98: Tenth European Conference On Machine Learning, 1998.

Hang Li and Kenji Yamanishi. Document classi_cation using a _nite mixture model. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, 1997.

Ray Liere and Prasad Tadepalli. Active learning with committees for text categorization. In AAAI-97, 1997.

Andrew McCallum, Ronald Rosenfeld, Tom Mitchell, and Andrew Ng. Improving text clasi_cation by shrinkage in a hierarchy of classes. In ICML-98, 1998.

Tom M. Mitchell. Machine Learning. WCB/McGraw-Hill, 1997.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Learning to classify text from labeled and unlabeled documents. In AAAI-98, 1998.

H. Heaps, A theory of relevance for automatic document classification, Information and Control, pp. 268–278, 1973