Study on Naive Bayesian Classifier and its relation to Information Gain

Anjana Kumari
Delhi Technological University
Delhi
anjana83.kumari@gmail.com

Abstract- Classification and clustering techniques in data mining are useful for a wide variety of real time applications dealing with large amount of data. Some of the application areas of data mining are text classification, medical diagnosis, intrusion detection systems etc. The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. The approach is called "naïve" because it assumes the independence between the various attribute values. Naïve Bayes classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and are then used to predict the class membership for a untrained data.

Keywords: *Classification, Bayesian Formula, Information Gain, weighted naive Bayesian classifier, Kullback's Test, Conditional Probability, Marginal Independence*

*****

I. INTRODUCTION

Classification techniques analyze and categorize the data into known classes. Each data sample is labeled with a known class label. Clustering is a process of grouping objects resulting into set of clusters such that similar objects are members of the same cluster and dissimilar objects belongs to different clusters. In classification the classes are pre-defined. Training sample data are used to create a model, where each training sample is assigned a predefined label. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Other than collection and managing data, data mining also includes analysis and prediction. In this report we will try to understand the logic behind Bayesian classification. The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

II. DATA MINING

Data Mining is the technology utilising the data analysis tools to operate on large amount of data to extract or discover previously unknown hidden pattern and relationships [5]. The data mining techniques can be categorized into one of the three methods which are association rule mining, classification & prediction and clustering.

III. CLASSIFICATION

Classification is a supervised learning technique that classifies data item into pre-defined class label. [6]This technique builds model that predict future data trend. There are several algorithms for data classification such as Decision Tree, CART (Classification and Regression Tree) and Back Propagation neural network. Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data. The model generated by a learning algorithm should both fit the input data well and correctly predict the class labels of records it has never seen before.

IV. BAYESIAN CLASSIFICATION

Bayesian Classifiers are statistical classifiers. Bayesian Classification is based on Bayes Theorem which utilises the conditional probability to classify the data into pre-determined classes[2]. The approach is called "naïve" because it assumes the independence between the various attribute values. Naïve Bayes classification can be viewed as both a descriptive and a predictive type of algorithm. The probabilities are descriptive and are then used to predict the class membership for a untrained data. The naïve Bayes approach has several advantages:

- It is easy to use.
- Only one scan of the training data is required.
- Easily handle mining value by simply omitting that probability.

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$$

- It requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.
- In spite of their naive design and apparently over-simplified assumptions, naive bayes classifiers

have worked quite well in many complex real world situations.

A. CONDITIONAL PROBABILITY & BAYES' RULE

Bayes classification has been proposed that is based on Bayes rule of conditional probability. Conditional Probability can be defined as: What is the probability that something will happen, given that something else has already happened.

Bayes rule is a technique to estimate the likelihood of a property given the set of data as evidence or input[2]. Let X be an input data sample whose class label is not known. Let H be any hypothesis, such that X belongs to a specified class C. For classification we need to determine P(H|X), the probability that the hypothesis H holds given the input data sample is X. P(H) is the priori probability of H, is the probability that the class label is C regardless of input data sample. In the same way , P(X|H) is posterior probability of X conditioned on H. P(X) is the prior probability of X. Bayes rule or Bayes theorem provides a way of calculating the posterior probability P(H|X), given P(H), P(X) and P(X|H).Formula is given by:

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \quad \text{Eq.1}$$

Bayes formula can be expressed informally in English by saying that

$$\text{Posterior} = \frac{\text{likelihood} \times prior}{evidence}$$

We know how frequently some particular evidence is observed, given a known outcome. We can use this known fact to compute the reverse, to compute the chance of that outcome happening, given the evidence.

B. ALGORITHM

The working of the Naive Bayes Classifier can be summarised as follows:[1]

1. Each of the input data item can be represented as an n-dimensional feature vector, such as
   X =(x_1,x_2,$x_{3,}$……….x_n).
2. Suppose that there are m classes, C_1,C_2,……..$C_{m.}$ Given a data item X whose class label is not known, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. Therefore the Naïve Bayes Classifier assigns an unknown data item to the class C_i if and only if
   $P(C_i|X) > P(C_j|X)$ for $1 \leq j \leq m$, $j \neq i$.

   By Bayes Theorem :
3. Only $P(X|C_i)$ $P(C_i)$ needs to be maximized as P(X) is constant for all classes. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, i.e., $P(C_1) = P(C_2)$ =…….= $P(C_m)$ and we would therefore maximise $P(X|C_i)$. Class prior probability can be estimated by $P(C_i) = S_i/S$, where S_i is the number of training samples of class C_i, and S is the total number of training samples.
4. If the data item has many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation, Naïve Bayes assumes that the attributes are conditionally independent of one another.

Thus, $P(X|C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$.

The probabilities $P(x_1|C_i)$,$P(x_2|C_i)$,$P(x_3|C_i)$, ….,$P(x_n|C_i)$ can be estimated from the training samples, where

(a) If A_k is categorical, then $P(x_k|C_i)=S_{ik}/S_i$, where S_{ik} is the number of training samples of class C_i having value x_k and S_i is the number of training samples belonging to C_i .

(b) If A_k is continuous-valued, then the attribute is assumed to have a Gaussian distribution so that

$$P(x_k \mid C_i) = g(x_k, \mu c_i, \sigma c_i) = \frac{1}{\sqrt{2\pi\sigma c_i}} e^{-\frac{(x_k - \mu c_i)^2}{2\sigma^2 c_i}}$$

Eq.2

Where $g(x_k, \mu c_i, \sigma c_i)$ is Guassian(normal) density function for attribute A_k , while µci and σci are the mean and standard deviation, respectively.

5. In order to classify an unknown data input X, $P(X|C_i)$ $P(C_i)$ is evaluated for each class C_i. X is then assigned to class C_i if $P(X|C_i)\, P(C_i) > P(X|C_j)\, P(C_j)$ for $1 \leq j \leq m$, $j \neq i$.

C. WEIGHTED NAIVE BAYESIAN

A new weighted Naive Bayesian Classifier model is proposed by DUAN Wei[8], which is based on information gain. According to which we can use information gain of an attribute to reduce attribute set, and assign relative weight to each classification attribute. In order to improve the classification efficiency propose a method of assign weight to attribute. The formula of weighted NBC as follow:

Argmax $P(h_j|x\ )$= argmax$\prod_{k=1}^{n} w_k P(x_k \mid h_j)$ P (h_j) ……. Eq.3

Where w_k represents the weight of attribute x_k .The larger the weight, the greater effect the attribute has on classification.

D. INFORMATION GAIN & ENTROPY

In information theory, entropy is a measure of the uncertainty in a random variable[9]. Entropy, H is a measure of impurity which is given by

$H(S)=-\sum p_i \log_2 p_i$ ……Eq.4

The measure of purity is called the information. It represents the expected amount of information that would be needed to specify whether a new instance should be classified as X or Y. Information gain is widely used method for calculating the importance for features in decision making.

Information_Gain = Entropy_before - Entropy_after

E. KULLBACK'S TEST FOR CONDITIONAL INDEPENDENCE

Information gain in terms of conditional probability[8] is given by I(X/Y) = H(X) - H(X|Y) where X,Y,Z are discrete random variables. Marginally Independent [8] term can be defined as Random variable X is marginally independent of random variable Y , knowledge of Y 's value doesn't affect your belief in the value of X for all xi ∈ dom(X), yj ∈ dom(Y) and yk ∈ dom(Y),

$$P(X = xi|Y = yj)$$

$$= P(X = xi|Y = yk)$$

$$= P(X = xi)$$

A test for marginal independence between X and Y can be set up in terms of information gain measures as follows:

$H_0 : \; I(X|Y) =0, (X \perp Y)$

$H_1 : I (X|Y) >0, (X \not\perp Y)$

A test for conditional independence of X and Y given Z, as proposed by Kullback, is based on the information gain on X given by Y once Z is known. The hypotheses to be tested are:

$H_0 : \; I(X|Z,Y) =0, (X \perp Y ,Z)$

$H_1 : \; I(X|Z,Y) >0, (X \not\perp Y ,Z)$

F. CONCLUSION

Applications of supervised learning are in almost any field or domain. The Naive Bayes algorithm affords fast, highly scalable model building and scoring. It scales linearly with the number of predictors and rows. The build process for Naive Bayes is parallelized. Naive Bayes can be used for both binary and multiclass classification problems. Naive Bayes handles missing values naturally as missing at random. The algorithm replaces sparse numerical data with zeros and sparse categorical data with zero vectors. Missing values in nested columns are interpreted as sparse. Missing values in columns with simple data types are interpreted as missing at random.

V. REFERENCES

[1] Jiawei Han,Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann.

[2] Richard O. Duda,Peter E. Hart,David Stork, "Pattern Classification", Second Edition,Wiley.

[3] Duan Weil,Lu Xiang-yang, "Weighted Naïve Bayesian model based on information gain", 2010 International conference on Intelligent System Design and Engineering Application, IEEE computer Society

[4] Chang-Hwan Lee, Fernando Gutierrez Dejing Dou , "Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure" 11th IEEE International Conference on Data Mining 2011.

[5] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[6] Jian Li, Jin-Mao Wei, Tian Yu and Hai-Wei Zhang ,"Feature Selection Based on Bayes Minimum Error Probability", 2012, 9th International Conference on Fuzzy Systems and Knowledge Discovery.

[7] http://en.wikipedia.org/wiki/Entropy_(information_theory)

[8] Manuel Martínez-Morales, Nicandro Cruz-Ramírez, José Luis Jiménez-Andrade, and Ramiro Garza-Domínguez, "Bayes-N: An Algorithm for Learning Bayesian Networks from Data Using Local Measures of Information Gain Applied to Classification Problems" MICAI 2004, LNAI 2972, pp. 527–535, 2004. Springer-Verlag Berlin Heidelberg