Edo Bander
Math 4500
Research Project

Bayesian Algorithm

Executive Summary

Bayesian learning algorithms are methods of calculating probabilities for hypotheses and are one of the most practical approaches to certain type of learning problems. Bayesian classifier is competitive with other learning algorithms in many cases and in some situations, it outperforms other methods. The reason why Bayesian algorithm is important to machine learning is because it provides a perspective for understanding other learning algorithms. Bayesian analysis is also used to justify a design choice in neural network learning algorithms.

This research analyzes a CardiologyCategorical.xls data set. This data set contains 302 different instances of 14 different attribute. These attributes are: age, sex, chest pain type, blood pressure, cholesterol, blood sugar, resting heart rate, angina, peak, slope, colored vessels, thal and class. The contribution of all attributes is independent and each contributes equally to the classification problem. By analyzing the contribution of each attribute, a conditional probability is determined. The classification that will be analyzed is made by combining the impact that different attributes have on prediction.

Given the data set, first we estimate the prior probability for each class by counting how many times each class appears in our data set. For each attribute, *a,* the number of occupancies of each value *a* can be counted to determines P (*a*). When we classify certain instance, the conditional and prior probability generated from data set is used. This is done by combining the different attribute values from the given instance. This would be known as a conditional probability or P (T | a). An example that we're going to use for instance is T = [M, 154] from our data set. This means that we're trying to classify in this situation a chest pain type from attributes Gender (male) and Heart rate. At this point, we have both prior and conditional probability that we needed. From these two, we calculate a likelihood that *a* is in each class. Class types in our example are: Asymptomatic, Abnormal Angina, Angina and No Tang. By adding all likelihoods, we calculated the estimated probability or P (T).

Bayesian theorem is a cornerstone of Bayesian learning methods because it calculates the final or posterior probability P (*a* | T), from prior probability P (*a*), together with estimated probability P (T) and conditional probability P (T | *a*):
Bayesian theorem:

$$P(a \mid T) = \frac{P(T \mid a) \times P(a)}{P(T)}$$

The results of the algorithm done with the instance T = [M , 154] are:

P (Asymptomatic) = 50%
P (Abnormal Angina) = 14%
P (Angina) = 8%
P (No Tang) = 28%

By looking at the results of the algorithm, we can classify instance T = [M , 154] as a Asymptomatic since it has highest probability out of four.

Problem Description

As we mentioned, this research analyzes a CardiologyCategorical.xls data set. This data set contains 302 different instances of 14 different attribute. These attributes are: age, sex, chest pain type, blood pressure, cholesterol, blood sugar, resting heart rate, angina, peak, slope, colored vessels, thal and class. Out of these 14 attributes, we selected few that will be compared in order to classify new instance. The contribution of all attributes is independent and each contributes equally to the classification problem. The two attributes that are mainly focused on were Gender and Heart rate. The classification that will be analyzed is made by combining the impact of these two attributes on a prediction of chest pain type. In order to do our classification, we divided Heart rate attribute into four ranges. These range are 0-100, 100-130, 130-170 and 170 to infinity. Before we started calculating our probabilities, two tables were created. One shows all the counts that we did for attribute Gender and each range of Heart rate attribute. This means what we counted how many instances of different chest pain type occurs in each of our attribute range. The second table shows probabilities that are based on our previous counts. From there, the Bayesian algorithm can be applied to the instance that we will try to classify.

Analysis Technique

Bayesian learning algorithm is a method that calculates probabilities for hypothesis. Bayesian algorithm is very important in understanding other machine learning algorithms. This method creates a hypothesis that makes a prediction in a probabilistic form.

In this research, Bayesian algorithm was applied to a CardiologyCategorica.xls data set. This is a data set that contains 302 different instances collected with 14 different attributes (age, sex, chest pain type, heart rate, cholesterol, blood pressure, blood sugar, resting, angina, peak, slope, colored vassals, thal, and class).

From the give data set, there are 142 instances of asymptomatic chest pain, 50 instances of abnormal angina, 87 instances of no tang and 23 instances of angina chest pain. The maximum heart rate attribute was divided into four ranges:

(0 – 100] , (100 – 130] , (130 – 170] , (170 – inf.]

With the given data, the prior probabilities were estimated. This was done by dividing number of instances of the certain chest pain type by the total number of instances. These are the results that we got:

P (Asymptomatic) = 142 / 302 = 0.4702 = 47%
P (Abnormal angina) = 50 / 302 = 0.1656 = 17%
P (Angina) = 23 / 302 = 0.0761 = 8%
P (No tang) = 87 / 302 = 0.2881 = 28%

By comparing three attributes from data set (chest pain type, gender, and maximum hart rate), we created a table that shows counts and probabilities with the chosen attributes.

Counts

		Count (Chest pain type)			
Attribute	Value	Asymptomatic	Abnormal Angina	No Tang	Angina
Gender	M	104	32	19	52
	F	38	18	4	35
Heart Rate	0 - 100	6	0	0	2
	100 - 130	39	3	3	6
	130 - 170	84	28	12	57
	170 - inf.	13	19	8	22

Probabilities

		Probabilities (Chest pain type)			
Attribute	Value	Asymptomatic	Abnormal Angina	No tang	Angina
Gender	M	104 / 142	32 / 50	19 / 23	52 / 87
	F	38 / 142	18 / 50	4 / 23	35 / 87
Heart Rate	0 - 100	6 / 142	0	0	2 / 87
	100 - 130	39 / 142	3 / 50	3 / 23	6 / 87
	130 - 170	84 / 142	28 / 50	12 / 23	57 / 87
	170 - inf.	13 / 142	19 / 50	8 / 23	22 / 87

The values from these two tables we can use to classify any tuple from the give data set. For example, we are going to classify
T = [M , 154]

By using these two values and calculated probabilities of gender and heart rate, the following estimates are calculated:

P (T | Asymptomatic) = 104/142 x 84/142 = 0.4332
P (T | Abnormal Angina) = 32/50 x 28/50 = 0.3584
P (T | Angina) = 19/23 x 12/23 = 0.4309
P (T | No Tang) = 52/87 x 57/87 = 0.3916

Combining these with prior probabilities, we estimate a likelihood of each chest pain type:

Likelihood of having Asymptomatic = 0.4702 x 0.4332 = 0.2037
Likelihood of having Abnormal Angina = 0.1656 x 0.3584 = 0.0594
Likelihood of having Angina = 0.0761 x 0.4309 = 0.0328
Likelihood of having No Tang = 0.2881 x 0.3916 = 0.1128

The estimated probability P(T) would be a sum of all likelihood values because T would be classified as either Asymptomatic, Abnormal Angina, Angina or No Tang:

P(T) = 0.2037 + 0.0594 + 0.0328 + 0.1128 = 0.4087

And finally, the actual probability of each chest pain type would be:

$$P(\text{Asymptomatic}) = \frac{0.4332 \times 0.4702}{0.4087} = 0.50 = 50\%$$

$$P(\text{Abnormal Angina}) = \frac{0.3584 \times 0.1656}{0.4087} = 0.15 = 14\%$$

$$P(\text{Angina}) = \frac{0.4309 \times 0.0761}{0.4087} = 0.08 = 8\%$$

$$P(\text{No Tang}) = \frac{0.3916 \times 0.2881}{0.4087} = 0.28 = 28\%$$

Based on these probabilities, T [M , 154] would be classified as Asymptomatic since it has highest probability of 50 %.

Assumptions

The contribution of all attributes is independent and each contributes equally to the classification problem. In this research, 14 different attributes could have been used. Out of all these attributes, few had to be selected in order to do classification. In order to classify chest pain type, we selected attributes gender and heart rate. Gender is selected because logically males and females have different reasons for chest pain. Heart rate was selected because we found that hart rate could be a reason for some chest pains. Assumption that had to be made is to divide hart rate into different ranges of value. Therefore, our results might have been different if these ranges were divided differently. The classification analyzed was made by combining the impact of these two attributes on the prediction of chest pain type.

Results

The results of the research will be presented in five steps. These are the steps:

1) First thing found was the prior probability of each class:
   P (Asymptomatic) = 0.4702 = 47%
   P (Abnormal angina) = 0.1656 = 17%
   P (Angina) = 0.0761 = 8%
   P (No tang) = 0.2881 = 28%

2) After the prior probability, conditional probability was estimated:
   P (T | Asymptomatic) = 0.4332
   P (T | Abnormal Angina) = 0.3584
   P (T | Angina) = 0.4309
   P (T | No Tang) = 0.3916

3) Combining the prior and conditional probability, likelihood of each class is estimated by multiplying the two:
   Likelihood of having Asymptomatic = 0.4702 x 0.4332 = 0.2037
   Likelihood of having Abnormal Angina = 0.1656 x 0.3584 = 0.0594
   Likelihood of having Angina = 0.0761 x 0.4309 = 0.0328
   Likelihood of having No Tang = 0.2881 x 0.3916 = 0.1128

4) Sum of all these likelihoods gave us an estimated probability:
   P(T) = 0.2037 + 0.0594 + 0.0328 + 0.1128 = 0.4087

5) And finally, actual probability was obtained:
   P(asymptomatic) = 50 %
   P(Abnormal Angina) = 14 %

P(Angina) = 8 %
P(No tang) = 28 %

And the conclusion of the classification tell that instance T = [M , 154] is classified as Asymptomatic because it has highest probability of 50 %.

Issues

The Bayesian learning algorithm has several advantages and disadvantages. Advantages of Bayesian algorithm is that is really easy to use and unlike other classification methods, it only requires on scan of training data. The naïve Bayes approach can easily handle missing values by simply omitting that probability when calculating the likelihoods of membership in each class.

Although the Bayesian algorithm is straightforward, it does not always give us results that are satisfied enough to do our classification. The attributes that we would use are not always independent. We could use subset of attributes by ignoring any that are dependent on others.

The attribute Heart rate, we divided into ranges of (0 – 100] , (100 – 130] , (130 – 170] , (170 – inf.] in order to solve our classification. Division of these ranges is not an easy task and how this is done can effect the results that we got.

Another practical difficulty in applying Bayesian algorithm is that they require knowledge of many probabilities. When the probabilities are not known, they are often estimated based on the background knowledge and previously available data.

Work cited

Dagum, P., & Luby, M. (1993). Approximating probabilistic reasoning in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 141 – 153.

Heckerman, D., Geiger, D., & Chickering, D. (1995) Learning Bayesian Networks: The combination of knowledge and statistical data. *Machine Learning*, 197. Kluwer Academic Publishers.

Margaret H. Dunham. First Edition. Data Mining: Introductory and Advanced Topics. Prentice Hall, 2004.