Alyssa Curran
May 1, 2008
Data-Mining Foundations

Final Project:
Bayesian Classification of Adolescent Self-Esteem Data

Executive Summary

The objective of this research project was to do an analysis on a large set of data collected by a psychological researcher. I used the method of Bayesian Classification to analyze a portion of this data. The original research paper from which I obtained the data was entitled: *Does adolescent self-esteem predict later life outcomes? A test of the causal role of self-esteem* (Boden 319). The question I endeavored to answer was whether or not a low measure of life satisfaction and whether or not the individual has anxiety disorder at age 18 predicts whether or not the individual's self-esteem was low at age 15. I used these two attributes and a "global measure from the Coopersmith Self Esteem Inventory" (Boden 322) as the labels to do an analysis with the Bayesian Classification algorithm. The algorithm requires that there only be two attributes, since it is based on probability. The probability of the individual being classified as a certain class (1-5 depending on the self-esteem inventory) depends on a specific combination of the probability of one attribute given the other, and the probability of the other attribute, given the first one (Dunham 87). Using a formula of conditional probabilities that I calculated from the data (using a training set to find probabilities, and a sample set to test the algorithm), I was able to classify the individuals in the sample data. The algorithm does require that the attributes be categorical. One of the attributes I used in the analysis was not originally categorical (life satisfaction), but I developed ranges to transform it into a categorical attribute.

I would like to add some background information about the subject of self-esteem. Many researchers "regard it as a form of evaluation of the self that guides future behavioral choice and action" (Boden 319). There have been links established between low self-esteem and a range of life outcomes: including substance abuse, mental illness, (including anxiety disorder), suicidal thoughts and behavior, and social problems. It has been determined that an individual's self-esteem is a critical determining factor of success and failure across a variety of life's undertakings.

I used 935 data entries as the training data set (life satisfaction quantitatively ranged from 12 to 40, and the anxiety data included either a 0 for no anxiety disorder, and a 1 for having the disorder), and I pulled out 50 data entries, 5%, to use as the test data to evaluate the algorithm after I calculated the necessary probabilities. When I tested the 50 data instances in the algorithm, **I obtained a 30% accuracy rate, which in my opinion is not sufficient to conclude that a low measure of life satisfaction and whether or not an individual has anxiety disorder at 18 predicts low self-esteem at age 15**. I used two measures of performance to get a better idea of how useful my results were. I used a Confusion Matrix – the best result would include zeros around the diagonal entries but my result includes non-zero numbers scattered throughout the matrix. I also used an Operating Characteristic Curve to evaluate the results. This type of curve plots the percentages of false positives on the x-axis against the percentages of true positives against the y-axis. The best result would show a fairly steeply sloped line close to the y-axis. My OC-Curve, however, shows a line that is quite flat, and is positioned near the x-axis. Based on these two performance measures and my own percentage of accuracy, my results are not significant enough to assume that the presence of anxiety disorder and life satisfaction at age 18 influence self-esteem at age 15.

Problem Description

The question I hope to answer in this analysis is whether or not low measures of life satisfaction and the presence of anxiety disorder at age 18 predict that an individual had low measures of self-esteem at age 15. This is a problem I wish to address using the Bayesian Classification technique, which uses conditional probabilities to classify test data (Dunham 86). I separated out 50 data entries from a total of 935 training entries (which is 5% of the data) to test the performance of my Bayesian Classification algorithm. The data I used for this analysis was obtained from a research project completed by Joseph M. Boden, David M. Fergusson, and L. John Horwood. The research paper was entitled: *Does adolescent self-esteem predict later life outcomes? A test of the causal role of self-esteem*. This data included many attributes and missing data. I eliminated the missing data entries, cut down most of the attributes to those I felt were sufficient to work with in this algorithm, and I ended up with a pre-processed portion of the data including two attributes (one categorical, one quantitative), and one labeling category. It is this sub-section of the data that I submitted to the algorithm, which uses probabilities that I calculated from the training data, to obtain my results.

Analysis Technique

I will outline here my methods for analyzing the data to answer my research question. I used the Bayesian Classification approach to extract results from the data. Before I go into the details of how the Bayesian Classification algorithm works, I would like to include some background information about the research paper from which I obtained my data. The research was done by Joseph M. Boden, David M. Fergusson, and L. John Horwood, who are from the Christchurch School of Medicine and Health Sciences. The paper studies the relationship between "self-esteem in adolescence and later mental health, substance use, and life and relationship outcomes in adulthood" (Boden 319). The study used data from a birth cohort of about 1,000 adults from New Zealand studied to the age of 25.

Self-esteem is seen by researchers as a "form of evaluation of the self that guides future behavioral choice and action" (319). They have established links between low self-esteem and a range of outcomes including mental illness, substance abuse, and social adjustment problems. Self-esteem is also seen as being critical in determining an individual's success and failure throughout an array of life's tasks. In the study, the adolescent age range was the focus of research because adolescence is thought to be a paramount time in the development of self-esteem. Events that take place in adolescence can have a great impact on later adult behaviors, including maladaptive behaviors in adolescence. It is thought that self-esteem can be implied from the development of certain adolescent behaviors. For example, resiliency or positive adaptation implies the development of high self-esteem, while maladaptive responses to certain issues that arise during adolescence imply the development of low self-esteem. This study was one of the first to study the relationship between self-esteem and life outcomes over a relatively long period of time. A longer time period of research can help to highlight certain relationships that may not be evident through short-term studies. The purpose of this study is to reveal insights as to whether or not low self-esteem is actually a cause of a range of disadvantageous life outcomes, so that if this is the case, efforts at raising self-esteem can be executed and will not be in vain. The study did include, however, a following of family background, and information on the individual's social and emotional context to allow for the controlling of covariate factors (which are those that cause low self-esteem *and* the specified later life outcomes, such as family environment). These contexts may be the cause of both low self-esteem *and* later life outcomes if the effects are reduced after controlling for covariates (Boden 319-322).

Now, I would like to transition to an explanation of the Bayesian Classification algorithm. This technique is based on Bayes' rule of conditional probability, and its usefulness depends on whether or not the contribution by all of the attributes are independent, and that each attribute contributes equally to the classification problem. If the attributes are not independent, this may significantly impact the results and render Bayesian Classification ineffectual. Using this method, a classification is made by combining the impact of each attribute on the prediction of a certain data instance. The approach is also called *naïve* Bayes Classification; naïve because it assumes that the attribute values are independent (Dunham 86-87).

First, I needed to pre-process the data so that it was useable. There were many missing entries in the data, so each data instance that was missing data (surprisingly, each instance that had missing data from the original dataset had a lot of missing attributes) was deleted from the database. I had to delete around 200 instances from the data to obtain a fully useable dataset. To begin my computations of the probabilities to use in the algorithm, I had to separate out a section of the data to use as test data, and the rest of the data was used to train the Bayesian algorithm. I sorted the data by the class (which was the score the individual received for self-esteem at 15; 1-5). I picked the first 10 data instances for each class, for a total of 50 test data instances – which is 5% of the data. The remaining data, 935 instances, was used to calculate the probabilities used in the algorithm.

Bayesian Classification stems from Bayes' original theorem. Bayes' theorem relates conditional probabilities (the probability of A given B) and prior probabilities (the separate probability of A) to obtain desired conditional probabilities (the probability of B given A). The theorem can be visualized as follows:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Using this theorem as a base, the Bayesian Classification algorithm attempts to provide a process in which to simply classify data instances based on data with known classifications. Using the conditional probability relation in Bayes' original theorem, it is possible to calculate the probabilities of the data instances occurring given they are a certain class, P(B|A), the probability of each class occurring in the data, P(A), and the probability of the data instance itself, P(B). Combining these pieces of the probability equation in a certain way will allow me to classify data instances into one of the five classes (Roiger 302).

To begin, it is important to first calculate the probability of each class, $P(C_i)$, occurring in the data. This is done by counting the occurrence of class 1 data instances and dividing it by the total number of instances, and doing the same for class 2, 3, 4, and 5. After that step is completed, it is necessary to find the probabilities of each value of each attribute. The Bayesian Classification algorithm only handles categorical data, so for the life satisfaction attribute, I divided it into ranges of values so that it was categorical. For this analysis, the necessary probabilities would be the probability of a data instance having anxiety at age 18, the probability of a data instance not having anxiety at age 18, and the probability of a data instance being in each range of the life satisfaction scores (one probability for (10-15], one for (15-20], (20-25], (25-30], (30-35], and (35-40]). These probabilities were calculated by summing the number of data entries that had the above attribute value, and dividing by the total number of data entries. It is important to note that I was able to count the data instances by sorting by the specific attribute I was focusing on (using Microsoft Excel).

There are a few more steps necessary to complete the probability table. I filled in a table which was labeled with the anxiety values (0 and 1) and the life satisfaction ranges (listed above) vertically, and horizontally, the table was labeled with the classes. I calculated the probabilities of each of the attribute values given they were from a given class. For example, I filled in the top row of the table with the probabilities of a data instance having a 0 value for anxiety (no anxiety disorder) *and* being from class 1, next the probability of the instance having a 0 for anxiety *and* being from class 2, etc. I did this for each attribute value, and divided the counts for each particular class by the number of data instances in that class. These probabilities allow us to obtain P(B|A) for each test data instance – the probability of a data instance occurring given it is from a certain class. Here is my table of probabilities:

		Probabilities				
Attribute:	**Value:**	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Class 5**
Anxiety Disorder	**0**	0.86096	0.77451	0.71779	0.65625	0.52047
	1	0.13904	0.22549	0.28221	0.34375	0.47953
Life Satisfaction	**(10-15]**	0.18717	0.12255	0.10429	0.05625	0.03509
	(15-20]	0.27272	0.2304	0.20245	0.21875	0.16374
	(20-25]	0.4492	0.57843	0.57055	0.55625	0.60819
	(25-30]	0.08556	0.06373	0.11656	0.1625	0.16374
	(30-35]	0.00535	0.0049	0.00613	0.00625	0.01754
	(35-40]	0	0	0	0	0.00585

The last step of the analysis involves actually performing the computations of the probabilities for each of the test data instances. To classify the test data, I first had to identify the values of each of the two attributes. Next, I marked down for each of the classes the value for that attribute value (according to my probability table). Each of the two probability values for each attribute value were then multiplied together to obtain P(B|A) – or the probability of that instance occurring given it is from a certain class. After multiplying those values, I end up with 5 values (one for each class) and then multiply each of those values by the probability of each class occurring in the data, or P(A). After this multiplication, the 5 values are then added together to obtain P(B), or the probability of the data instance occurring itself. Lastly, to obtain the final probabilities of the data instance being in a certain class, I divide each of the values obtained by the summed value. This will give me the desired probability, or P(A|B), for each class. To actually classify the data instance, the probability with the highest value will be assigned as the class. Below is an example of my calculations for a test data instance:

Data attribute values: Anxiety (18) = 0, Life Satisfaction (18) = 16

P(t|1) = .86096 x .27272 = .23480 x P(C_1) = .049613321

P(t|2) = .77451 x .23040 = .17845 x P(C_2) = .041133570

P(t|3) = .71779 x .20245 = .14532 x P(C_3) = .026764524

P(t|4) = .65625 x .21875 = .14355 x P(C_4) = .025953390

P(t|5) = .52047 x .16374 = .08522 x P(C_5) = .016466577

Summed: P(t) = .159931382

P(1|t) = .049613321/.159931382 = **.3102**

P(2|t) = .041133570/.159931382 = .2572

P(3|t) = .026764524/.159931382 = .1673

P(4|t) = .025953390/.159931382 = .1623

P(5|t) = .016466577/.159931382 = .1030

*This highest value is P(1|t), so this data instance is classified as class 1. This is actually an accurate classification for this particular data instance.

Assumptions

1) The data is accurate.
2) The data is still useful after some data has been omitted due to missing values.
3) The data attributes (variables) are independent of each other.
4) The method picked is sufficient to evaluate the data.
5) 5% of the data extracted from the dataset is sufficient to use for test data.

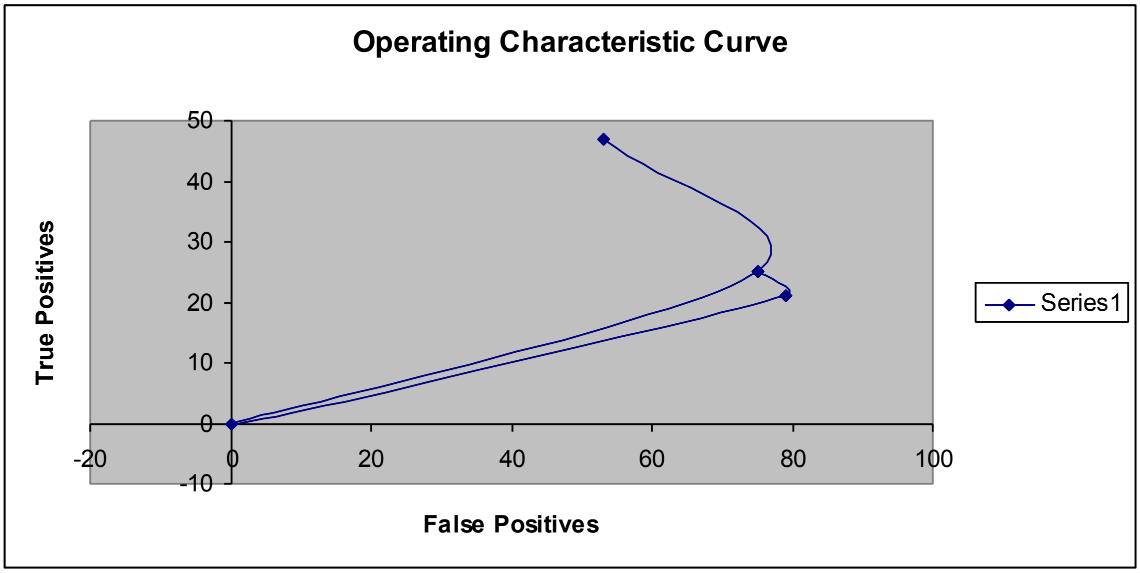
Results

After I performed the complete Bayesian Classification analysis on the 50 test data instances, **I obtained a 30% accuracy rate of classification.** This, in my opinion, is not high enough to assume that there is a significant relationship between anxiety disorder at 18, life satisfaction at 18, and self-esteem at 15. **I, therefore, cannot conclude that a low measure of life satisfaction and whether or not an individual has anxiety**

disorder at age 18 predicts that they had low self-esteem at age 15. To facilitate a visualization of my results and to further illustrate them, I will provide here two performance measures of my analysis. The first one is a *Confusion Matrix*, which is a table showing the relationship between the actual class of the data entry and the class assigned by the algorithm. Visually, the best solution would show mostly zeros around the diagonal (I have highlighted the diagonal). Mine shows many non-zero numbers scattered throughout the matrix, and hence my low accuracy rate. The matrix looks as follows:

	Assigned Class				
Actual Class	**1**	**2**	**3**	**4**	**5**
1	3	7	0	0	0
2	1	4	0	2	3
3	5	2	0	0	3
4	2	5	0	1	2
5	1	1	0	1	7

The second performance measure is the Operating Characteristic Curve. The OC-Curve measures the false positives (classified as a specified class when it is not actually in that class) on the x-axis against the true positives (specified accurately as a specific class) on the y-axis. The better the results, the more steep the line will be, and the closer to the y-axis. My results, however, show the line to be closer to the x-axis and rather flat. The OC-Curve looks as follows:



Issues

With Bayesian Classification, it is important to keep in mind that the best results are obtained if the data attributes are independent of each other and if they contribute equally to the problem. In the original research paper from which I obtained the data, the researcher states that there are many covariates that can be accounted for, and that some of the data attributes have been found to be dependent on each other. My analysis of the data confirms in a much more simplified way the researcher's results.

However, I would like to mention that my analysis is different from the researcher's in that it was done in the opposite direction. The researcher took an assessment of many individuals' self-esteem at age 15 to evaluate whether or not low self-esteem scores at this age preceded later life outcomes, such as anxiety disorder and low life satisfaction, among others. I, however, conducted my analysis from the viewpoint that the presence of anxiety disorder at age 18 and low life satisfaction scores at age 18 predict that the individual had low self-esteem at age 15. It is a situation of reversed cause and effect. Both analyses imply that there are covariates and that the attributes are dependent on each other to an extent.

Appendices

I have no appendices to add here.

References

Boden, Joseph M.; Fergusson, David M.; Horwood, L. John. (2008). Does adolescent self-esteem predict later life outcomes? A test of the causal role of self-esteem. *Development and Psychology*, 20, 319-339.

Dunham, Margaret H. (2003). *Data Mining: Introductory and Advanced Topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Roiger, Richard J.; Geatz, Michael W. (2003). *Data Mining: A Tutorial-Based Primer*. Boston, MA: Pearson Education, Inc.

* The raw data was obtained directly from Joseph M. Boden.