Predicting the Probability of Having an X-Linked Recessive Inheritance Using Bayes' Theorem

Arif Çakır

Mathematical Engineering- Science and Letters Faculty

090190355

Mat 116E, CRN: 20662

Doç. Dr. Burcu Tunga

June 6, 2021

Predicting the Probability of Having an X-Linked Recessive Inheritance Using Bayes' Theorem

Abstract: Bayes' theorem (or Bayes' rule) is a theorem that describes the probability of an event based on prior knowledge of the conditions that might be related to the event. It was named after and found by English mathematician Thomas Bayes (1701(?)-1761) and published by his friend Richard Price (1723-1791) after two years after his death. According to Bayes' theorem, when A and B are events and the probability of event B (*P(B)*) is not equal to zero, the probability of event A occurring (*P(A)*) given that event B has occurred can be showed as *P(A|B)* and equal to $\frac{P(B|A)P(A)}{P(B)}$. If P(A) probability is taken as the probability of gonosomal (genetic characters that linked to the chromosome X) recessive (genetic characteristics that expressed in offspring only when inherited from both parents) inheritance, and B is a condition that A requires, P(A|B) conditional probability can be calculated with the help of Bayes' theorem.

In this report, the main goals are predicting the status of each person in a family relative to a gonosomal recessive inheritance, comparing predicted and original values, and visualizing the performance of the predictions. To visualize the performance of the predictions, a confusion matrix is used.

Introduction:

Bayes' Theorem has applications in scientific fields that are not directly related to mathematics such as sociology, archaeology[1], or medicine. Two of these fields are genetics and genetic counseling. Genetic counseling is the process of advising people affected by or at risk of genetic disorders. According to Demirel and Bodur (2004), Bayes' theorem can be used in genetic counseling for calculating the risk of recurrences of genetic inheritances, by providing the use of all suitable information together (p. 84). One kind of these genetic inheritances is gonosomal recessive inheritances.

Gonosomal Recessive Inheritances:

Gonosomal recessive inheritances are inheritances that are linked to chromosome X and not express in the phenotype of offspring unless inherited from both parents. Colorblindness, hemophilia A and B, and Duchenne muscular dystrophy are the most popular examples of gonosomal recessive inheritances. Due to females having two X chromosomes and males having one X and Y chromosome, females can carry the gene of inheritance without expressing it in phenotype, when males have to express the gene of inheritance in phenotype if they have it. A male's status relative to inheritances is independent to their father and dependent on their mother, while a female's status relative to the same inheritance is dependent on both their father and mother. Because of this, gonosomal recessive inheritances seen more in males than females and females can have the genes of inheritance without expressing it in phenotype.

[1] Otarola-Castillo E. Toruqato M. G. *Bayesian Statistics in Archaeology* (Annual Reviews, July 26, 2018)

Classes and Class Labels:

Bayes classification is based on estimating the probability P(X|Y) of predictor X and class Y. According to the Mathworks website, it has two steps:

1. Training step: Estimating the parameters of a probability assuming that predictors are independent given to the class, by using the training data.
2. Prediction Step: Computing the posterior probability of the sample for any test data, then classifying the test data according to the largest posterior probability.

In this project, variables were chosen as sex of the person (Named as “Sex” with the labels of {‘M’} for male and {‘F’} for female), generation of the person (Named as “Gen”, with integer numbers between 1 and latest generation for the label), Status of the mother of the person relative to inheritance (Named as “Mother”, with the labels of {‘Yes’} for affected, {‘No’} for normal, and {‘Carrier’} for carrying the inheritance), Status of the father of the person relative to inheritance (Named as “Father”, with the labels of {‘Yes’} for affected and {‘No’} for normal) for predicting the variable “Person”. Labels of the “Person” variable are {‘Yes’} if the person is affected by inheritance, {‘No’} if the person is not affected by inheritance, and {‘Carrier’} if the person is female and carry the genes of the inheritance without showing it in phenotype. To test these variables, two sample families were used for both test and prediction.

Chosen families:

Due to the possibility of mutations and difficulty of obtaining correct information for all members of a real family, two families that do not exist are used for testing: a randomly generated family and a hand-made family. The Hand-made family is named “Made family” and the randomly generated family is named “Random Family” for file and program names. The number of members of the Random Family and their status generated with the help of the “randomfamily.m” program which can be found in the homework file. Made Family was created for testing the case of preservation of inheritance in each generation, and Random Family was made for testing a randomly made family. The families visualized with a pedigree tree and that pedigree tree transformed into a table as a .xlsx file. As Allen and Darwiche highlighted, it is best to representing genotype for all members in the pedigree for the gene that modeled (2007, p.505). Because of that, pedigree trees display the status of the person as genotype and their sex. Both pedigree trees and .xlsx files can be found in the homework file, pedigree trees were not displayed in the report because of not fitting to the page.

Training:

Training made with the “bayes” function which can be found in the homework file. “bayes” function trains the data regardless of the input family with the help of built-in MATLAB function “fitcnb”, but input must be a table and contain variables “Sex”, “Gen”, “Mother”, “Father”, and “Person”. After testing the data, the “bayes.m” function uses “individual” or “family” functions for prediction, depending on user input.

Prediction:

- Predicting a single individual:

  Prediction of a single individual made by “individual” function that can be found in the homework file. “individual” function creates a 1x4 table “ind” for an individual with the classes of “Sex”, “Gen”, “Mother”, “Father” in the input table of “bayes” function and then predicts the status of the person relative to the gonosomal recessive inheritance with the help of built-in MATLAB function “predict”. “predict” function takes table “ind” and trained classification from “bayes” function “mdl” and predicts the status of the individual relative to inheritance.

- Predicting the whole family:

  Prediction of whole family made by “family” function that can be found in the homework file. “family” function creates the variable “label” for each member of the family with the help of the inputs of classification “mdl” from “bayes” function and table “t”, the input table of function “bayes”. Later, “family” creates a cell array named “perf” and tests the performance of the prediction with the help of MATLAB built-in function “strcmp”. After that, “family” creates a table named “t3” that contains the variable names “Names”, “Sex”, “Generation”, “Original”, “Predicted”, and “Performance”. While variables “Names”, “Sex”, “Generation”, and “Original” are the same as variables “Name”, “Sex”, “Gen”, and “Person” from the input table of function “bayes”, “Predicted” is the predicted value which is contained in variable “label” and “Performance” is the array “perf”. Lastly, the “family” function displays table “t3” with the help of built-in

MATLAB function “uitable” or confusion matrix of classes “Original” and “Predicted” from table “t3” with row and column summaries.

Testing:

- Made Family:

  Firstly, when looked at the confusion matrix in Figure 1, The individuals whose true class is “carrier” predicted as “carrier” with %100 accuracies, but 4 out of 19 people whose true class is “no” predicted as “carrier” and 4 out of 18 people whose true class is “yes” predicted as “carrier” which makes it %21.05 and %22.22 in order. The main reason for these errors is the high population of individuals with the “carrier” class in the data set. In addition to that, row summaries and column summaries show that 9 out of 19 individuals whose real class is “no”, this makes it %47,4. The main reasons for this error are the shortness of the data set (56 individuals), and more importantly, the distribution of chromosomes given to offspring being uneven. Moreover, according to the “MadePerformance.xlsx” file in the homework file, no males were predicted as a “carrier” due to the absence of the “carrier” males in both Made Family and the real world. Finally, the status of 43 individuals from Made Family with 58 members has been predicted right, which means the prediction made with 25.86% error.

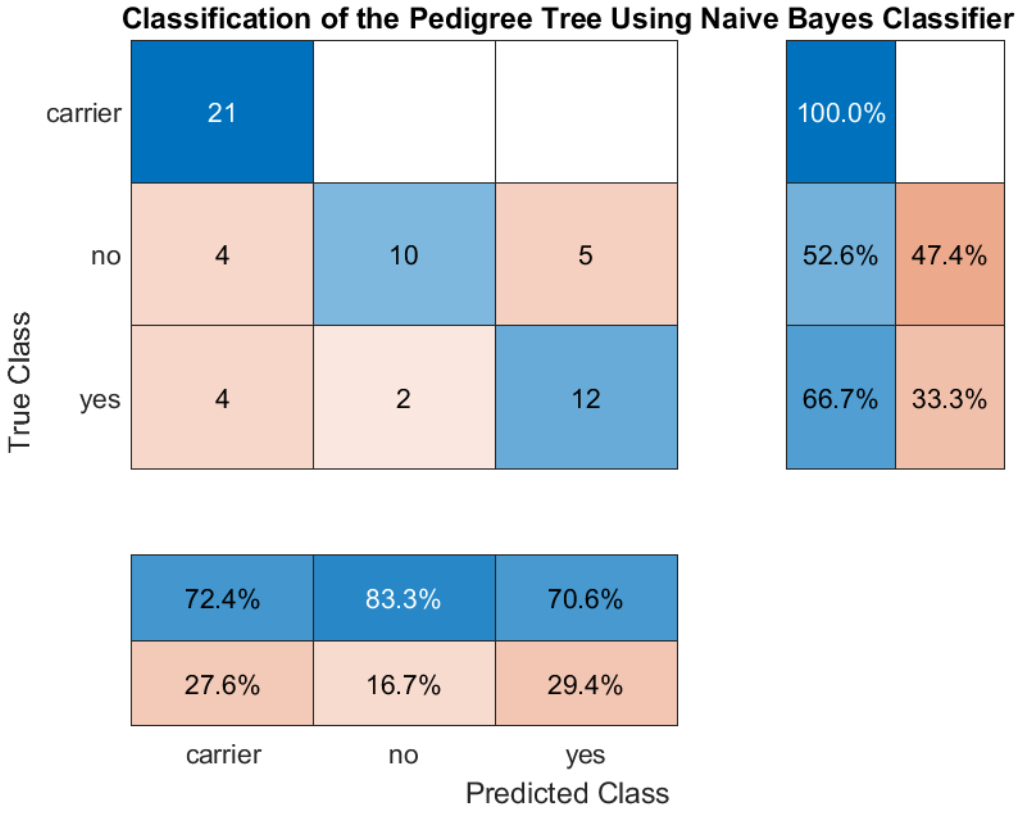

Figure 1: Confusion matrix of Made Family

- Random Family

Secondly, when the confusion matrix in Figure 2 is examined, it can be seen easily that predictions about Random Family were more accurate than Made Family. Random Family predicted with 89.8% accuracy which means the whole family except 5 individuals predicted right. No individual with "M" label predicted as "carrier" just as Made Family and even though they form only 18.37% of the data, people with "yes" label predicted with 88.9% accuracy. Furthermore, unlike Made Family, "carrier" and "no" labels predicted with 92.3% and 88.9% accuracy respectively, despite being close to each other. Finally, the status of 44 individuals from Random Family that formed by 49 members predicted right, which means prediction made with 10.2% accuracy.

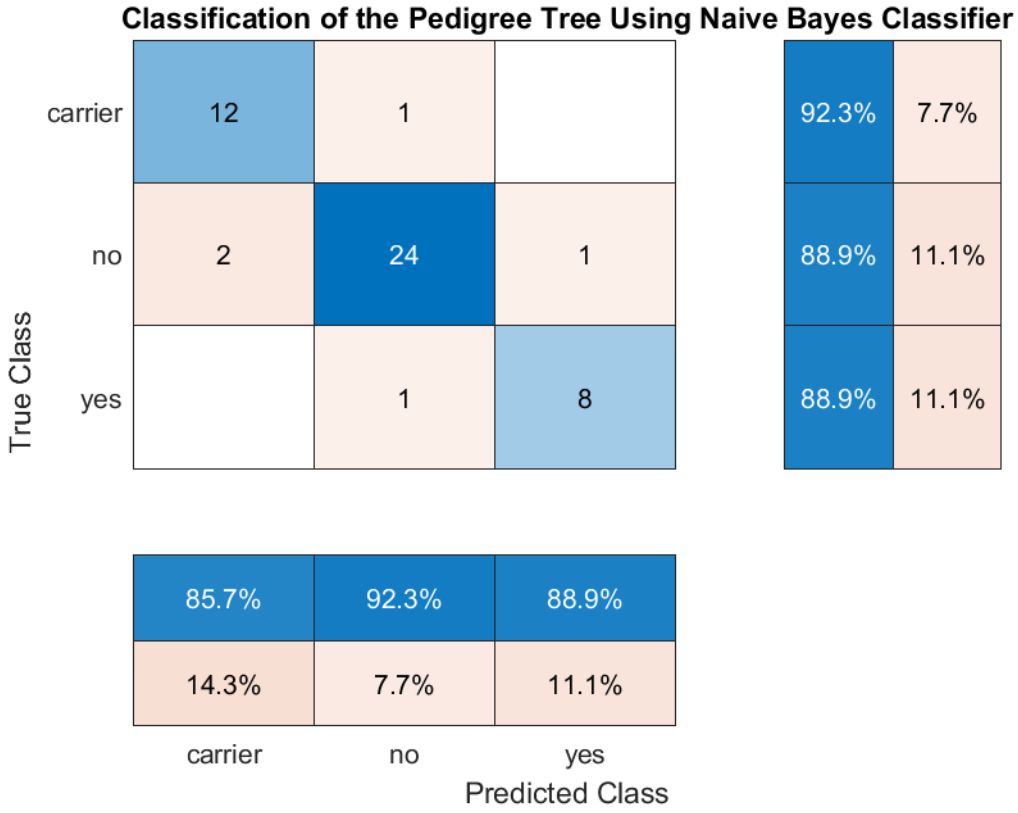

Figure 2: Confusion matrix of Random Family

Study on Test Data:

After testing data sets Made Family and Random Family with errors of 25.86% and 10.2% respectively, it can easily be seen that random data is more accurate than hand-made one. Because of being created for testing the preservation of tested inheritance, Made Family contained 21 “carrier” females and 18 individuals with class “yes”, which are 4 females and 14 males. Due to 67.24% of Made Family having or at least carrying the inheritance, 47.4% of people who do not show the inheritance in their genotype predicted wrong. This wrong prediction occurred because of the small size of individuals without inheritance compared to individuals with inheritance. For example, when looked at F16, their labels are ‘F’ for Sex, 4 for Gen, ‘carrier’ for Mother, ‘no’ for Father. F16

was predicted as a "carrier" despite being a member of the "no" class. When Bayes' Theorem applied on F16 as Y = (Person = "carrier"),X = (Person = "no"), and

T = (Sex = 'F', Gen = 4, Mother = 'carrier', Father = 'no'), the probability of F16 being a carrier is equal to $\frac{P(T|Y)P(Y)}{P(T)}$ and probability of F16 not having the inheritance is $\frac{P(T|X)P(X)}{P(T)}$.

Calculations for F16:

P(Person = "carrier") = P(Y) = 0.3621, P(Sex = 'F' | Y) = 1, P(Gen = 4| Y) = 0.619, P(Mother = 'carrier' | Y) = 0.333, P(Father = 'no' | Y) = 0.5714,

P(T|Y) = 0.1178.

P(Person = "no") = P(X) = 0.3276, P(Sex = 'F' | X) = 0.2105, P(Gen = 4 | X) = 0.5263,

P(Mother = 'carrier' | X) = 0.5789, P(Father = 'no' | X) = 0.5789,

P(T|X) = 0.0122.

P(T) = 0.0883.

When Bayes' theorem used with calculated probabilities, P(Y | T) = $\frac{0.1178x0.3621}{0.0883}$ = 0.4831 and P(X | T) = $\frac{0.0122x0.3276}{0.0883}$ = 0.0453. This means that the probability of F16 being a "carrier" is 0.4831 while the probability of F16 being a "no" is 0.0453. Due to this high difference between the two probabilities, F16 was predicted as a "carrier" despite being a member of the class "no". It can be seen from this example that all wrong predictions, just like F16, caused by the high amount of "carrier" class in the data set.

Furthermore, when the Random Family data set is checked, there are 27 “no”, 13 “carrier” and 9 “yes”. Even though label “no” forms more than half of the data set, it did not affect the prediction of the other two labels as the “carrier” label did in Made Family. The reason for this difference is places of individuals with “no” and “carrier” labels in both families. While individuals in Made Family were placed for the preservation of inheritance and variation, individuals in Random Family were placed randomly to be more realistic.

Conclusion:

In conclusion, the status of individuals in two families that does not exist predicted and these predictions tested in this project. Hand-made family predicted with a high error such as 25.86% while randomly generated family predicted with 10.2% error. For improving the project, more independent labels can be added and new randomly generated families with more members can be tested. Also, with the replacement of the “Gen” label with another label, predicting the status of next generations with Bayes’ Theorem may be possible and even can be used in genetic counseling.

References

Allen D. & Darwiche A. (2007). RC_Link: Genetic linkage analysis using Bayesian networks. International Journal of Approximate Reasoning, pp. 499-523.

Demirel, S. & Bodur, S. (2004). Application of Bayes Theorem in genetic counseling. Erciyes Tıp Dergisi, pp. 81-85.

Naïve Bayes Classification, Retrieved from Mathworks website: https://www.mathworks.com/help/stats/naive-bayes-classification.html