Ayush Thada

Bayesian Statistics: Techniques & Models

Final Report

Abstract:

The fertility potential analysis represents a research topic of great interest and it could help us to understand and examine all the factors which affect the fertility and the extents up to which fertility is affected. There are several studies which has studied this effect using the non-Bayesian supervised and unsupervised methods. In this study we try to analyze the factors affecting the fertility in the males using the Bayesian framework and try to provide some good insights about the issue.

Introduction:

The World Health Organization (WHO) defines infertility as a disease. It's a reproductive system disease that manifest as the inability of obtain clinical pregnancy after 12 months of having unprotected sexual relationships. Infertility causes several effects in different types of personal health: physical, mental, emotional, psychological, social and even religious, in the couples that suffer from it. It's one of the most important causes of depression, and its social, psychological and cultural consequences have been catalogued in six levels of severity, ranging from guilt, fear, depression to violent death or suicide [1].

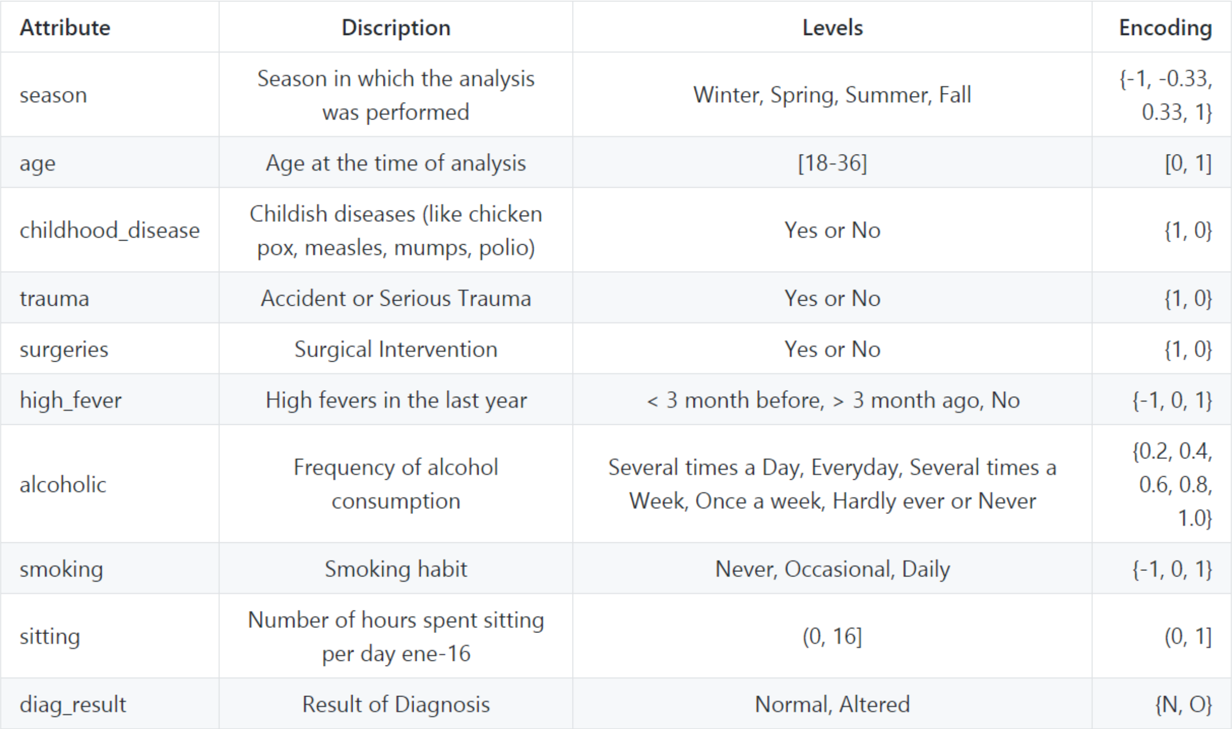
This behavior is also confirmed by [2], and clearly infertility can affect negatively generating frustration, and personality weakness. The fertility potential analysis represents a research topic of great interest, and could help us to understand all the factors that difficult to have high fertility rates, which is quite relevant to be able to propose solutions and obtain an increase in the fertility levels especially for men. Some authors [3], say that during the last three decades, several reports have suggested that the semen quality in regular men has decreased, and [4] talks about a trend in decrease of sperm count and seminal fluid volume in the last fifty years.

In [5] they mention that fertility rates have decreased drastically in the last two decades, especially in men, due to environmental issues and lifestyle that can affect the quality of semen. Several artificial intelligence techniques have become an emergent technology for decision support systems in medicine to patient identification with fertility problems.

In [6] a research was conducted to study semen volume, sperm concentration, progressive motility, vigor and percentage of normal forms and multiple anomalies. The semen volume didn't decrease, but an important decrease in total sperm count was found (443,2 million in 1976 to 300.2 million in 2009), also in motility (64% in 1976 to 49% in 2009) and vigor (88% to 80%). In [7] the authors consider that semen analysis is standard for routine diagnosis of infertile couples studies through the sperm count, and it´s strongly related to male infertility, and they also express the importance of sperm concentration in male infertility, since it is a relevant factor in diagnosis of this disease.

Data:

For this study, we're using the "Fertility Dataset" [8], which is available at UCI machine learning dataset repository. Dataset has 10 attributes and 100 instances. The descriptions as well as the other details are available in the following image, available at my Github repository [9].



Attribute	Discription	Levels	Encoding
season	Season in which the analysis was performed	Winter, Spring, Summer, Fall	{-1, -0.33, 0.33, 1}
age	Age at the time of analysis	[18-36]	[0, 1]
childhood_disease	Childish diseases (like chicken pox, measles, mumps, polio)	Yes or No	{1, 0}
trauma	Accident or Serious Trauma	Yes or No	{1, 0}
surgeries	Surgical Intervention	Yes or No	{1, 0}
high_fever	High fevers in the last year	< 3 month before, > 3 month ago, No	{-1, 0, 1}
alcoholic	Frequency of alcohol consumption	Several times a Day, Everyday, Several times a Week, Once a week, Hardly ever or Never	{0.2, 0.4, 0.6, 0.8, 1.0}
smoking	Smoking habit	Never, Occasional, Daily	{-1, 0, 1}
sitting	Number of hours spent sitting per day ene-16	(0, 16]	(0, 1]
diag_result	Result of Diagnosis	Normal, Altered	{N, O}

During the exploratory data analysis, it has been found that there is huge class invariance, 88% of the records has normal result of diagnosis. For instance if we consider all attributes as numerical variables, we obtain there is not very strong correlation between the variables. The absolute correlation values lies between 0 and 0.27. Therefore dimensionality reduction will lead to high information loss in this case. But to reduce the model complexity in some cases like Cell Mean model, feature selection can be considered. For this experiment, we're assuming that the samples are independent and identically distributed. Also we're working using Bayesian framework hence it's implicit that we're assuming that data is fixed and parameters of the process through which data is generated is random variables.

Model:

The problem which is undertaken is a binary classification problem. Hence the appropriate distribution for this case is Bernoulli distribution. It'll produce the values between [0, 1] that can be considered as the probabilities for an instance to belong to a particular class. But still two model choices are there which is discussed in the next paragraph. One important thing to note here is that for both of models Laplacian prior or double exponential priors have been used. All priors provide some degree of regularization that is, helping parameters avoid straying from sensible regions. On a log-scale, Gaussian priors say that distant regions become less likely as a quadratic function of their distance; Laplace priors say that this fall-off is linear, that is, much slower. As such, Laplace priors have "thicker tails" and penalize distance less, asymptotically, than Gaussian priors. In simple language, Laplace priors say being far away is less problematic than Gaussian priors do. For the same reason standard Student-T distribution can be used as well.

A. Additive Model

In this model, all the categorical variables' categories are converted into an identifier whose values are either 0 or 1. Then linear combinations of all these variables are taken and corresponding to which probability is estimated from a Bernoulli distribution. Here variables of linear combination are parameter of the model and the target of the model is to estimate the distribution of these variables. Here no of parameters are

$$|\ \mathbf{N}\ |\ + \sum_{i=1}^{n} |\ \mathbf{Ci}\ |$$

where | N |=> Count of attributes with numerical values

| C_i | => Number of categories in i^{th} categorical variable.

B. Cell Mean Model

In this model, all the dataset is converted into categories, because this model is most suitable for discrete categorical values. Here both categorical variable 'age' and 'sitting' is discretized into 5 categories. Here every object of the cross product of all the categories in considered as a random variable and the target is to estimate the distribution of that random variables. Here no of parameters are

$$\Pi_{i=1}^{n} |\boldsymbol{Ci}\ |$$

where | C_i | => Number of categories in i^{th} categorical variable.

Results:

Conclusions:

References:

[1] Villalobos, A. Centro de Especialidades Ginecológicas y Obstetricas. Recuperado de http://infertilidadcr.com/publicaciones/infertilidadpubli.html

[2] Brugo-Olmedo, S., Chillik, C., & Kopelman, S. (2003). Definición y causas de la infertilidad. Revista colombiana de Obstetricia y Ginecología, 54, 227-248.

[3] Auger, J., Kunstmann, J. M., Czyglik, F., & Jouannet, P. (1995). Decline in semen quality among fertile men in Paris during the past 20 years. New England Journal of Medicine, 332(5), 281-285.

[4] Carlsen, E., Giwercman, A., Keiding, N., & Skakkebæk, N. E. (1992). Evidence for decreasing quality of semen during past 50 years. Bmj, 305(6854), 609-613.

[5] Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. Expert Systems with Applications, 39(16), 12564-12573.

[6] Splingart, C., Frapsauce, C., Veau, S., Barthelemy, C., Royère, D., & Guérif, F. (2012). Semen variation in a population of fertile donors: evaluation in a French centre over a 34-year period. International journal of andrology, 35(3), 467-474.

[7] Bonde, J. P. E., Ernst, E., Jensen, T. K., Hjollund, N. H. I., Kolstad, H., Scheike, T., ... & Olsen, J. (1998). Relation between semen quality and fertility: a population-based study of 430 first-pregnancy planners. The Lancet, 352(9135), 1172-1177.

[8] Fertility Data Set, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Fertility

[9] Project Code and notebook Repository, https://github.com/itsayushthada/Bayesian-Statistics/tree/master/Project-01