# Bayesian Kernel Methods:
## Applications in Medical Diagnosis Decision-Making Processes (A Case Study)

Arti Saxena, Manav Rachna International Institute of Research and Studies, Faridabad, India

https://orcid.org/0000-0002-4162-793X

Vijay Kumar, Manav Rachna International Institute of Research and Studies, Faridabad, India

## ABSTRACT

In the healthcare industry, sources look after different customers with diverse diseases and complications. Thus, at the source, a great amount of data in all aspects like status of the patients, behaviour of the diseases, etc. are collected, and now it becomes the job of the practitioner at source to use the available data for diagnosing the diseases accurately and then prescribe the relevant treatment. Machine learning techniques are useful to deal with large datasets, with an aim to produce meaningful information from the raw information for the purpose of decision making. The inharmonious behavior of the data is the motivation behind the development of new tools and demonstrates the available information to some meaningful information for decision making. As per the literature, healthcare of patients can be analyzed through machine learning tools, and henceforth, in the article, a Bayesian kernel method for medical decision-making problems has been discussed, which suits the purpose of researchers in the enhancement of their research in the domain of medical decision making.

## KEYWORDS

Bayesian Inference, Bayesian Kernel, Bayesian Network Structure, Kernel Mean, Kernels, Medical Decision, Non-Parametric Bayesian Kernel Method

## INTRODUCTION

Many people all around the world die due to error in healthcare systems. In healthcare industry, several strategies have been proposed by various professionals like IT adoption, collaboration among various disciplines etc. to design medical decision support systems, which help clinicians for medical decision making. From the last three decades, the applications of Bayesian approaches have grown at an exponential pace, but research in this domain has developed very slowly in first one hand half decade. The major reason for that is to generate Bayesian networks for practical and analytical purposes are quite difficult. Therefore, generating Bayesian networks was a challenge for researchers and this made inaccessible to vast community of scientist for their applicability. Now, Bayesian and its native forms are generated by the help of computers, which address the vast community of researchers to became a tool and have wings spread over many disciplines such as computer science, logic, information theory, probability theory, statistics, machine learning etc. and can be utilized in almost all the disciplines for the purpose of application. Bayesian based models are mainly used to answer the queries for the variables, their relationship to compute the evidence for the variables in complex situations and are considered as graphical probabilistic models. Decisions making under the state of

uncertainty are the typical most problems in medical decision making and Bayesian kernel methods are very promising in tackling such situations problems and becomes more rational as compared to the conventional models, the assumptions made are more obvious and thus making the decision making easy and defendable. The analytical approach of the theory became a boon for the researchers, which flavoured with machine learning algorithms to grow beyond their domain. The traditional statistical techniques such as regressions have explanatory purposes and have applications in simulation and optimization, epidemiology and social sciences. The foundation of Bayesian approach lies from the rule for updating probabilities, given by Rev. Thomas Bayes (1702-1761) as Bayes' theorem. He addresses both discrete and continuous probability distributions by means of conditional and marginal probabilities. Bayes' theorem is used for the calculation of inverse probability and is based on the apriori and posterior probability measures and provides uncontroversial result in the field of probability but its applications fall under controversy for more than two decades. This chapter is an introduction to the modernized Bayesian approach which leads to algorithms within the frame work of risk minimization and giving us a new insight into kernel algorithms. There are many ways we could take to motivate the readers for using Bayesian kernel methods in medical diagnosis, which is based on the assumption that all the quantities of interest are governed by probability distributions. Optimized results can be obtained in Bayesian theory by evaluating these probabilities together and combines prior knowledge with the observed data and is the sole theory in medical diagnosis which accommodates hypothesis.

## LITERATURE REVIEW

During the clinical examination of a patient, some patient-centric information is recorded such as age, gender, medical history and some parameters related to the disease. Clinicians prefer an interpretable decision support system, based on clinical and pathological indicators. Many researchers used SVM for this purpose, but due to the heterogeneous nature of clinical data, SVM is not easy to implement and the results obtained are not reliable. Therefore, to get better results, kernel methods should be implemented. Schurmann (1996) explained the concepts behind the designing of kernels and its properties. Ghosh (2000) elaborated that the Bayesian approach has been recognized as an promising technique for tackling clinical decision making problems and has the ability to represent uncertain knowledge. He represented Bayesian based mathematical model of heart disease. Spiegelhalter (2000) presented decision theoretic statistical based bayesian methods and its implementation in the assessment of health technology. He claimed that bayesian methods are the best in transforming the problem from initial opinion to final judgement. Scholkopf (2003) described the basic principles of Gaussian processes and their implementation in his chapter on Bayesian Kernel Methods. Sheppard (2005) presented a mathematical approach flavoured with Bayesian theory to measure the level of uncertainty and further utilize the assessment to propose the improved diagnosis. Due to which, one could treat the probability of false detection or missed detection and provide the treatment for false diagnosis. Kadane (2005) discussed the use of Bayesian approach in medical decision making and emphasized on how decision makers consider the subjective concepts of probability and utility functions in his research. Kim (2006) used Bayesian based Gaussian Process classifier kernel method to classify gender among various face images of men and women. The proposed method is efficient over the traditional support vector machines based kernel classifiers and found that they determine hyper parameters on model selection criterion. Van Calster (2007) used Bayesian least square support vector machines method to separate malignant from benign and develop a classifier to predict malignancy in adnexal masses. He used a large dataset collected from the nine databases of different centres and comes out with the better results as compared with the traditional support vector machine method. Broemeling (2007) explained various benefits of employing Bayesian methods in clinical studies and emphasizes that variety of areas where diagnostic medicine is used such as in the estimation of accuracy by sensitivity, positive and negative predictive values of diagnostic measurements. Lukic (2007) also

used Bayesian kernel method for analyzing functional neuroimages. Li (2008) developed an integrated approach by the help of real time data to identify the risk factors associated with occupational history of patients. He used Bayesian approach to process the data set by integrating medical knowledge representation and genetic algorithms. Penny (2011) used Bayesian methods in brain imaging. Kenzi (2012) explained the gram matrix representation of nonparametric Bayesian kernel methods. YUE (2012) performed a Meta analysis of neuroimaging data consisting of peak activation locations in 162 separate studies on emotions by nonparametric binary regression Bayesian method on the data of neuroimaging. Aram (2013) developed quantitative model based on kernel Bayesian approach to calculate the risk involved in osteolysis after hip replacement. Barbini (2013) in one of his chapter on Bayesian approach in medicine and health management explained the importance of Bayesian methods in medicine and healthy management as it makes the problems easier to explain and defend. Sidong (2013) proposed a Multifold Bayesian kernel based method to model the diagnosis process and the planning of the treatment for Alzheimer's disease, which overcome the constraint of the performance of machine learner with the multi-modal data. Molly (2014) analyzed the performance of Bayesian kernel methods over the traditional statistical method to estimate and forecast data. He used Poisson Bayesian kernel model against Poisson generalized linear model and found that Bayesian kernel approach tends to outperform classical data count models for smaller datasets, results have been validated through measures of goodness of fit and logarithm based likelihood function. Deck (2014) proposed Bayesian based kernel methods to improve results in classification and regression tasks in natural language processing problems. He used Gaussian processes over commonly used support vector machines and concluded that Bayesian based Gaussian approach fit the model easily even in case of complex kernel problems, SVM based techniques lacks in model selection, which hinders the use of advanced kernels. Page (2015) proposed a model for predicting health events from electronic health records using machine learning. Jessica (2015) proposed a two-stage method with naïve Bayes' kernel machine method to estimate gene set in first stage and regularizations with kernel principal component analysis stage two for the estimation of risk classification based on gene structure to overcome the limitation to address non-linearity in the existing methods. Jiang (2018) proposed constraint based learning algorithm for the construction of Bayesian networks, which has the advantage over the traditional methods of calculating correlation between the variables but accuracy of these existing methods are not high. He proposes entropy estimation based Gaussian kernel density estimator to improve the learning of concrete as well as continuous data with sparsity. Bayesian theory got recognition in 2011, when Prof. Judea Pearl received the prestigious Turing award for the creation of Bayesian networks, since then the theory of Bayesian networks got more public attention and became a probabilistic model of complex situations and a primary algorithm for inference in these models. This is an important tool to address multi disciplines and particularly revolutionized the area of artificial intelligence. Bayesian network has been recognized as a powerful tool to reasoning of uncertainty knowledge and widely used in the domains of medical diagnosis and prediction, fault diagnosis, financial analysis and prediction, decision support etc. Concerning about the data on the model source, Bayesian networks can be build from human knowledge, i.e through theory, or data. Bayesian networks are a versatile modelling framework, making them suitable for many problem domains. Predictive modelling techniques are very much popular and are algorithmic in nature and prediction of any disease can be viewed as a classification problem, as the outcome is a binary variable. Classification techniques can be applied on structured as well as unstructured data sets. The data collection may be homogeneous as well as heterogeneous. Homogeneity brings the benefits of certainty in the individual patient's need and accordingly resources are utilized. In this way, classification techniques provide the opportunity for improved clinical diagnosis and efficient planning of healthcare resources.

## BASICS OF BAYES'

In many situations, the practitioners consider some set of candidate hypotheses $H$ and are interested in finding the most probable hypothesis $h \in H$ for given observed set of data $\bar{D}$ and such hypothesis is called maximum a posterior (MAP) hypothesis. This is usually denoted as $h_{MAP}$ and can easily be determined by Bayes Theorem to find posterior probability of each candidate hypothesis i.e.:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \, P\left(\bar{D} \mid h\right) \cdot P\left(h\right) \tag{1}$$

In machine learning problems, Bayes' Theorem is introduced by considering the data $\bar{D}$ as training examples of target function and $H$ is domain of Candidate target function. And it has been noticed that this method is very successful in the sense that no other hypothesis is more likely.

To illustrate Bayes' Rule, case study of medical decision making problem has been considered by taking two alternative hypothesis viz as that the patient has a particular form of breast cancer and that the patient does not, according to an available laboratory test with two possible outcomes as: + (positive) and – (negative).

In case of Breast cancer, the classification of data can be considered by assigning probabilities to a particular type of Breast cancer $B_i \left\{i = 1, 2, --, n\right\}$ among $n$ classes given two observations $\bar{b}$, denoted by $P\left(B = B_i \mid \bar{b}\right)$.

By Bayes' Theorem:

$$P\left(B = B_i \mid \bar{b}\right) = \frac{p\left(B = B_i \mid \bar{b}\right)}{p\left(\bar{b}\right)} = \frac{p\left(\bar{b} \mid B = B_i\right) \cdot P\left(B = B_i\right)}{p\left(\bar{b}\right)} \tag{2}$$

where:

- $P\left(B = B_i\right)$ is the prior probability of Breast cancer of class i;
- $P\left(B = B_i \mid \bar{b}\right)$ is the posterior probability of Breast cancer given observation $\bar{b}$;
- $p\left(\bar{b} \mid B = B_i\right)$ is the conditional probability of observation $\bar{b}$ given class $B_i$; and
- $p\left(\bar{b}\right)$ is the probability of observation $\bar{b}$.

For an $n$ class classification probability:

$$P\left(B = B_i \mid \bar{b}\right) = \frac{p\left(\bar{b} \mid B_i\right) \cdot P\left(B_i\right)}{\sum_{N=1}^{n} p\left(\bar{b} \mid B = B_N\right) \cdot P\left(B = B_N\right)} \tag{3}$$

To find the inverse probabilities, hypothetical numerical data has been taken for the better understanding of the concept of Bayesian Inference. We have prior knowledge that among the entire population of people 0.015 have breast cancer of type $B_1$. Furthermore, the laboratory test is only an imperfect disease indicator. The test returns a correct positive result in only 95% of the cases in which the disease is actually present and the correct negative result in only 94% of the cases in which the disease is not present. In rest of the cases the lab test results in an opposite result.

The above data can be represented as follows:

$$P\left(B_1\right) = 0.015$$
$$P\left(+ \mid B_1\right) = 0.95$$
$$P\left(- \mid \sim B_1\right) = 0.94$$

From these following are the direct implications by probability sum rule as follows:

$$P\left(\sim B_1\right) = 1 - P\left(B_1\right) = 1 - 0.015 = 0.985$$
$$P\left(- \mid B_1\right) = 1 - P\left(+ \mid B_1\right) = 1 - 0.95 = 0.05$$
$$P\left(+ \mid \sim B_1\right) = 1 - P\left(- \mid \sim B_1\right) = 1 - 0.94 = 0.06$$

Now let us suppose that we have to observe a new patient and the labtest is positive for that particular patient. Now the question arises that, should we diagnose the patient as having Breast cancer of type $B_1$.

The maximum a posterior hypothesis can be found using Equation (1) as follows:

$$P\left(+ \mid B_1\right) \cdot P\left(B_1\right) = \left(0.95\right)\left(0.015\right) = 0.01425$$
$$P\left(+ \mid \sim B_1\right) \cdot P\left(\sim B_1\right) = \left(0.06\right)\left(0.985\right) = 0.05910$$

Thus combining the above expressions with Equation (1), we get:

$$h_{MAP} = \sim B_1$$

and the exact posterior probability can also be determined by normalizing the above quantities i.e. using Equation (3):

$$P\left(B = B_1 \mid +\right) = \frac{P\left(+ \mid B_1\right) \cdot P\left(B_1\right)}{P\left(+ \mid B_1\right) \cdot P\left(B_1\right) + P\left(+ \mid \sim B_1\right) \cdot P\left(\sim B_1\right)}$$

$$= \frac{0.01425}{0.01425 + 0.05910} = 0.1943 \left(approx.\right)$$

In the similar way:

$$P\left(B = B_2 \mid +\right), \; P\left(B = B_3 \mid +\right), \; P\left(B = B_i \mid +\right)$$

can be calculated and the most probable hypothesis for the type of Breast cancer can be determined by the following expression:

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \, P\left(B_i \mid +\right)$$

Thus to apply Bayes' theorem we need $P\left(+\right)$ directly although it's not given directly, we can calculate it as we know $P\left(B_1 \mid +\right)$ and $P\left(\sim B_1 \mid +\right)$ must sum to 1. This means either the patient has Breast cancer of type $B_1$ or does not. Here point is to be noted that the posterior probability of Breast cancer of type $B_1$ (0.1943), is higher than its prior probability (0.015), the most probable hypothesis is still that the patient does not have Breast cancer of type $B_1$. By means of this, Bayesian inference is strongly connected with the prior probability and applicable directly. Another point to be noted is that none of the hypothesis in Bayesian approach is completely rejected or accepted rather gives more or less probable values as cited in the above example. Thus to apply Equation (3), apriori and the probability density function of the given data is required. From the large data samples, the estimation of the apriori probability is obtained from the known class.

Kernel methods play a pivot role in decision making processes. With the inference of machine learning models flavoured with kernel models, gives efficient results and makes decision making a cake walk. From the past one decade, medical decision making developed as a domain on the account of large data, parallaly learning with kernels is a powerful tool to support decision making and have the capability to address large data in medical diagnosis. In machine learning problems, kernel is introduced as a similarity measure that can be found as a dot product in feature space.

## MATERIALS AND METHODS

### Parametric and Non-Parametric

Kernel based algorithms are very flexible in the sense of accommodation of domain specific prior information and predefined set of algorithms for implementation and Bayesian methods act as boon to the perceptive presentation of function spaces used by kernel methods. With the concern towards predictive modelling, Bayesian kernel method is very much promising, simple and easily integrable into decision making algorithms and used to calculate factors, which are more responsible for the diseases. Bayesian kernel methods enable the importance of factors in measuring probability distribution for their importance measure using data and prior available information to represent prior and conditional probability. The transformation from Bayesian to Bayesian Kernel approach requires some rules and estimates, kernel mean is one among those inferences. By this, "nonparametric" Bayesian inference is possible. It is a complete nonparametric approach where all calculations are done by linear algebra with Gram matrices.

### Parametric Tests

These tests take into account the assumptions about the population distribution parameters. So these tests rely on a fixed set of parameters.

### Non-Parametric Tests

These tests do not take any assumptions about the population distribution parameters. Thus, these tests are also called distribution free tests.

### Learning Bayesian Network Structure

Learning the structure of Bayesian networks, two algorithms are generally used such as constraint based algorithms and score based algorithms. The former is based on probabilistic semantic of Bayesian network and are responsible to identify marginal and conditional indecencies of the structure while

the laterer is based on the metric, which measures the quality of the candidate network by means of observed data. Bayesian networks are nonparametric in nature, so no hypothesis is formed to establish the relationships between variables, but through machine learning, an array of highly optimized learning algorithms are used to uncover structures in datasets. In Bayesian networks framework, diagnosis, prediction and simulation are computable and the inference from effect to cause and cause to effect. It gives access to expert system to work interactively in an automated process and became a bridge between AI and human intelligence.

In real life situations, unstructured or fragmented data or no data is available for computation, therefore it is required to build a robust model for decision support system. Bayesian networks are used to handle uncertainty with inputs, relationships and computed outputs as probabilistic and represent the entire probability distribution system under study. Naïve Bayes structure is commonly used network in which only one target (parent) node is connected to a set of nodes rather than computing the relationship between each node individually. Bayesian approach generates structures and parameters automatically from data. For example, in tumour classification problem, the medical diagnosis is performed by medical knowledge representation and artificial intelligence techniques to support doctors but by the intervene of Bayesian based methods the accuracy is over 95%. Unsupervised Structural Learning is the heuristic technique to form knowledge discovery as it puts no constrain on hypotheses for the exploration of relationship between the variables. In Bayesian beliefs offers variety of options to design and use algorithms in machine learning environment. The aim of implementing Bayesian techniques to learning algorithms is of risk minimization, which allows us to understand kernel algorithms in relatively different prospective and referred as Bayesian estimation. The kernel mean presentation of prior probability and conditional probability of Bayes' theorem has been discussed in next section.

## KERNEL METHODS - KERNEL MEAN

### Kernel Mean: Representing Probability

Classical Non-Parametric Approach

Kernel estimate for probability:

$$\hat{p}(x) = \frac{1}{n}\sum_{j=1}^{n} k\left(\left(x - X_j\right) / h\right)$$

Characteristic Function:

$$f_X(s) = E\left[e^{isX}\right] = \frac{1}{n}\sum_{j=1}^{n} e^{isX_j}$$

*New Alternative Non-Parametric Kernel Estimate*

Let $X$ be a random variable with probability $P$ by taking the values on $\bar{D}$, $k$ be a positive definite kernel on $\bar{D}$, and $H_k$ represents hypothesis associated with $k$, then the Kernel mean of $X$ on the hypothesis $H_k$ is defined as:

$$M_P \Rightarrow E\left[\psi(X)\right] = \int k(\cdot, x) dP(x) \epsilon H_k$$

Here $\psi\left(X\right) = k\left(\cdot, x\right)$ is a feature vector:

$$\hat{M}_P = \frac{1}{n}\sum_{j=1}^{n}\psi\left(X_j\right)$$

A bounded measurable kernel $k$ is said to be a characteristic kernel if:

$$M_P = M_Q \Rightarrow P = Q$$

i.e. a Kernel mean $M_P$ with characteristic kernel $k$ determines the probability uniquely.
Note:

1. $f_X\left(s\right)$ determines the probability of the random variable $X$ uniquely.
2. $k$ gives better insight of the alternative.
3. Kernels compute the measures effectively.
4. It can be applied to non-vectorial data also.

With $k$, any inference on $P$ implies an inference on $M_P$. So, our aim is estimating the kernel mean of the posterior probability given kernel knowledge representation of prior and conditional probability.

### Conventional Non-Parametric Approach

Kernel estimate according to conventional non-parametric approach:

$$\hat{p}\left(x\right) = \frac{1}{nh^d}\sum_{j=1}^{n}k\left(\left(x - X_j\right)/h\right)$$

Here $k\left(.\right)$ is chosen in such a way that it represents a symmetric probability distribution.
For example:

$$k\left(\bar{x}\right) = \left(2\pi\right)^{-\frac{d}{2}} exp\left(-\frac{1}{2}\bar{x}^T\bar{x}\right)$$

where $h$ explains the bias and variance in the estimates, hence taking the responsibility of the effectiveness of the model and is very important factor. Small values of $h$ causes random variations in density estimates and large values of $h$ can change or eliminate the underlying characteristics of distribution like bimodality. Also, $h$ is referred as smoothing parameter or window width or band width and provides the smoothness of density estimation which affects the posterior probability. $n$ is the number of observations $X_j$ with the dimension $d$. The range of $x$ depends on the sample data.

For example, for a normally distributed one dimensional data: $h = \left(\dfrac{4}{3n}\right)^{\frac{1}{5}} \sigma$ where $\sigma$ is standard deviation of data and the corresponding function is $E\left[e^{isX}\right]$.

## Kernel Mean: Representing Conditional Probability

Conditional probability is an important factor describing the conditional dependence and independence. The kernel estimate for conditional probability is defined as:

$$E\left[\psi\left(Y\right) \mid X = x\right] = \int \psi\left(Y\right) p\left(y \mid x\right) dy$$

To estimate kernel, exact estimation of $p\left(y \mid x\right)$ is not easy. To overcome such situation, we follow regression approach as the following.

Let $\left(X, Y\right)$ be the random pair taking values from $\overline{D_X} \times \overline{D_Y}$ .

$\left(H_X, k_X\right)$, $\left(H_Y, k_Y\right)$ be the corresponding hypothesis on $\overline{D_X}$ and $\overline{D_Y}$ respectively.

The covariance operators:

$$C_{YX} : H_X \rightarrow H_Y$$

$$C_{XX} : H_X \rightarrow H_X$$

where:

$$C_{YX} = E\left[\psi_Y\left(X\right) \psi_X\left(Y\right)^T\right]$$

and:

$C_{YX}$ can be determined by kernel mean $E\left[k_Y\left(.,Y\right) \otimes k_X\left(.,X\right)\right]$ on the product space $H_Y \otimes H_X$ .

## Kernel Bayes' Rule

Sum rule: $P\left(x\right) = \int p\left(x \mid y\right) \Pi\left(y\right) dy$
Chain Rule: $P\left(x, y\right) = p\left(x \mid y\right) \Pi\left(y\right)$
Bayes' Rule: $P\left(y \mid x\right) = \dfrac{P\left(x \mid y\right) \Pi\left(y\right)}{\int p\left(x \mid y\right) \Pi\left(y\right) dy}$

Algorithm for Kernelization:

**Step 1:** To express probabilities by means of kernel means.
**Step 2:** To express statistical inference rules with covariance operators.
**Step 3:** The gram matrix is computed by means of inference rules.

Now we will discuss kernel rules with Gram Matrix Computations in next section.

## GRAM MATRIX COMPUTATIONS

In this section, the Kernel methods approach with nonparametric tests using Gram matrix approach of linear algebra has been described.

## Kernel Sum Rule

Sum Rule: $P(x) = \int p(x \mid y) \Pi(y) dy$

Kernelization: $m_X = C_{XY} C_{YY}^{-1} m_\pi$

Gram Matrix representation: $\widehat{m_\pi} = \sum_{j=1}^{l} \alpha_j \varphi\left(\widehat{Y_j}\right), (X_1, Y_1), - - - (X_l, Y_l) \sim P_{XY}$

$$\widehat{m_X} = \sum_{j=1}^{l} \beta_j \varphi\left(\widehat{X_j}\right), \beta = \left(G_Y + n \in_n I_n\right)^{-1} G_{Y\tilde{Y}} \alpha, G_Y = \left(k\left(Y_i, Y_j\right)\right)_{ij}, G_{Y\tilde{Y}} = \left(k\left(Y_i, \widetilde{Y_j}\right)\right)_{ij}$$

Kernel Chain Rule

Chain Rule: $P(x, y) = p(x \mid y) \Pi(y)$

Kernelization: $m_X = C_{(XY)Y} C_{YY}^{-1} m_\pi$

Gram Matrix representation:

$$\widehat{m_X} = \sum_{j=1}^{l} \beta_j \varphi\left(X_j\right) \otimes \varphi\left(Y_j\right), \beta = \left(G_Y + n \in_n I_n\right)^{-1} G_{Y\tilde{Y}} \alpha,$$
$$G_Y = \left(k\left(Y_i, Y_j\right)\right)_{ij},$$
$$G_{Y\tilde{Y}} = \left(k\left(Y_i, \widetilde{Y_j}\right)\right)_{ij}$$

Note:

$$C_{(XY)Y} : H_Y \to H_X \otimes H_Y, E\left[\left(\varphi\left(X\right) \otimes \varphi\left(Y\right)\right) \otimes \varphi\left(Y\right)\right]$$

## Kernel Bayes' Rule

Bayes' Rule is regression $y \to x$ with the probability:

$$P(x, y) = P(x \mid y) \Pi(y)$$

According to kernel Bayes' Rule:

$$m_{P_{y|x}} = C_{YX}^\pi C_{XX}^{\pi -1} \varphi\left(Y\right)$$

where:

$$C_{YX}^{\pi} = C_{(YX)X} C_{XX}^{-1} m_{\pi}$$

and:

$$C_{YY}^{\pi} = C_{(YY)X} C_{XX}^{-1} m_{\pi}$$

Note: Mean on the product space is same as calculating the Covariance.

## BAYESIAN INFERENCE IN AI FOUNDATION

Bayesian inference methods can be used in analyzing network variables in AI-powered network systems and is considered as a powerful tool for modelling random variables of both discrete as well as continuous. These models use data as evidence and map the problems to evaluate data in the domain of probability. The methodology of these Bayesian network are analogous to that of Bayes' rule as discussed in the beginning of the manuscript, follows the following steps as:

1. Assign an initial prior probability distribution, which quantify the information into one distribution;
2. Choose a probabilistic model along with parameters, which relates the random variables;
3. Apply Bayes' theorem to use the prior knowledge base and the observed information obtained to get the posterior probability distribution.

This posterior distribution is re-configured according to the prior knowledge base and the observed data. In other words, we can say that "*Today's posterior is tomorrow's prior!*"

This methodology inspires to explore the Bayesian inference for AI systems.

## RESULTS AND DISCUSSION

Kernel mean is a non-parametric way of computing Bayes' rule and no parametric models are required to determine the probability inference rules. Many researchers have studied various properties of Bayesian estimators for three different types of prior probabilities defined as:

1. The adjacent coefficients are correlated by considering associated probabilities and termed as Gaussian processes;
2. The estimates of sparse matrix are expanded by means of linear combination of kernel functions and are referred as Laplacian processes;
3. Each kernel is governed by means of normal distribution with its mean and variance and is termed as vector machines.

Among these, only Gaussian and Laplacian kernels determines the probability uniquely and unable to determine Polynomial kernels.

## REFERENCES

Abrahamsen, P. (1992). *A Review of Gaussian random fields and correlation functions.* Rapport 917, Norwegian Computing Center.

Alexander, J. & Bernhard, S. (2003). Bayesian Kernel Methods- Advanced Lectures on Machine Learning. *LNAI, 2600*, 65–117.

Aram, P., Kandirkamanathan, V., & Wilkinson, J. M. (2013). Use of kernel-based Bayesian models to predict late osteolysis after hip replacement. *Journal of the Royal Society, Interface*, *2013*(Sept), 20130678. doi:10.1098/rsif.2013.0678 PMID:24047875

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, *68*(3), 337–404. doi:10.1090/S0002-9947-1950-0051437-7

Barbini E., Manzi P., Barbini P. (2013). *Bayesian approach in medicine and health management*. 10.5772/52402

Bashkirov, O. A., Braverman, E. M., & Muchnik, I. B. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, *25*, 629–663.

Bayes, T. (n.d.). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. (1763). Philosophical Transactions, Giving Some Account of the Present Undertakings. *Studies and Labours of the Ingenious in Many Considerable Parts of the World*, *53*, 370–418.

Beck, D. (2014). Bayesian Kernel Methods for Natural Language Processing. *Proceedings of the ACL 2014 Student Research Workshop,* 1-9. doi:10.3115/v1/P14-3001

Berg, C., Christensen, J. P. R., & Ressel, P. (1984). *Harmonic Analysis on Semigroups*. Springer-Verlag. doi:10.1007/978-1-4612-1128-0

Boser, B. E., Guyon, I. M., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In D. Haussler, (Ed.), *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, (pp. 44–152). ACM Press. doi:10.1145/130385.130401

Broemeling, L. D. (2007). *Bayesian biostatistics and diagnostic medicine*. Chapman & Hall/CRC. doi:10.1201/9781584887683

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. doi:10.1023/A:1009715923555

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. J. Wiley and Sons.

Cristianini, N., & Taylor, J. S. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Daemen, A., & Timmerman, D. (n.d.). Improved modeling of clinical data with kernel methods. *Elsevier Editorial System (tm) for artificial intelligence in medicine*, AIIM-D-09-00188R5.

Ghosh, J. K., & Valtorta, M. (2000). Building a Bayesian network model of heart disease. In *Proceedings of the 38th annual on Southeast regional conference, ACM-SE00 ACM-SE 2000- 38th Annual ACM Southeast Regional Conference*. ACM. doi:10.1145/1127716.1127770

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Computation*, *10*(6), 1455–1480. doi:10.1162/089976698300017269 PMID:9698353

Guyon, I., Boser, B., & Vapnik, V. (1993). Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, & C. Lee Giles (Eds.), Advances in Neural Information Processing Systems (Vol. 5, pp. 147–155). Morgan Kaufmann.

Haussler, D. (1999). *Convolutional kernels on discrete structures.* Technical Report UCSC-CRL-99-10, Computer Science Department, University of California at Santa Cruz.

Jaakkola, T. S., & Haussler, D. (1999). Probabilistic kernel regression models. *Proceedings of the 1999 Conference on AI and Statistics*.

Jessica, M., Ming, Y., & Jun, S. (2015). Risk Classification with an Adaptive Naive Bayes Kernel Machine Model. *Journal of the American Statistical Association*, *110*(509), 393–404. doi:10.1080/01621459.2014.908778 PMID:26236061

Jiang, L., & Zizhen, G. (2018). An improved constraint-based Bayesian network learning method using Gaussian kernel probability density estimator. *Expert Systems with Applications*, *113*, 544–554. doi:10.1016/j.eswa.2018.06.058

Kadane, J. B. (2005). Bayesian methods for health-related decision making. *Statistics in Medicine*, *24*(4), 563–567. doi:10.1002/sim.2036 PMID:15678444

Kenzi. (2012). Kernel Bayes' Rule: Nonparametric Bayesian inference with kernels. *NIPS 2012 Workshop Confluence between Kernel Methods and Graphical Models*.

Kim, H.-C. K., Daijin, G., Zoubin, B., & Sung Y. (2006). Gender Classification with Bayesian Kernel Methods. In *Proceeding of 2006 International Joint Conference on Neural Networks.* IEEE Explorer.

Kumar, V., Saxena, A., & Kumar, S. (2018). *Kernel based machine learning techniques: An innovative way of designing healthcare systems and services. In Big Data Management and the Internet of Things for improved Health systems*. IGI Global.

Li, J., Shi, J., & Satz, D. (2008). Modeling and analysis of disease and risk factors through learning Bayesian networks from observational data. *Quality and Reliability Engineering International*, *24*(3), 291–302. doi:10.1002/qre.893

Lukic, A. S., Wernick, M. N., Tzikas, D. G., Chen, X., Likas, A., Galatsanos, N. P., Yang, Y., Zhao, F., & Strother, S. C. (2007). Bayesian kernel methods for analysis of functional neuroimages. *IEEE Transactions on Medical Imaging*, *26*(12), 1613–1624. doi:10.1109/TMI.2007.896934 PMID:18092732

MacKay, D. J. C. (1998). Introduction to Gaussian processes. In C. M. Bishop (Ed.), *Neural Networks and Machine Learning* (pp. 133–165). Springer-Verlag.

Molly Stam, F., Hiba, B., & Kash, B. (2014). Empirical analysis of Bayesian kernel methods for modeling count data. In *Proceeding of the 2014 Systems and Information Engineering Design Symposium*, (pp. 25-25). IEEE Explore.

Page, D. (2015). *Predicting health events from electronic health records using machine learning. Talk by David Page at the Center for Predictive Computational Phenotyping*. University of Wisconsin.

Parmigiani, G. (2002). Modeling in medical decision making: A Bayesian approach. Academic Press.

Parthasarathy, K. R., & Schmidt, K. (1972). Positive definite kernels, continuous tensor products, and central limit theorems of probability theory. Lecture Notes in Mathematics, 272.

Penny, W. (2011). Bayesian methods in brain imaging, First Technical Course, European Centre for soft computing, Mieres, Spain.

Saitoh, S. (1988). *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical.

Sarika, J., Kumar, V., & Saxena, A. (2018). Generalized Fuzzy Information Entropy Measure: A Case Study for the Selection of Diamond among Various Brands. *Recent Patents on Engineering*, *12*(3), 223–229. doi:10.2174/1872212112666180301143158

Scholkopf, B. (1997). *Support Vector Learning*. R. Oldenbourg Verlag.

Scholkopf, B., Smola, A. J. & Muller, K.R. (1999). *Kernel principal component analysis*. Academic Press.

Schurmann, J. (1996). *Pattern Classification: a unified view of statistical and neural approaches*. Wiley.

Scholkopf, B., Smola, A., & Smola, J. (2002). *Learning with Kernels (Support Vector Machines, Regularization, Optimization, and Beyond)*. The MIT Press.

Sheppard, J. W., & Kaufman, M. A. (2005). A Bayesian approach to diagnosis and prognosis using built-in test. *IEEE Transactions on Instrumentation and Measurement*, *54*(3), 1003–1018. doi:10.1109/TIM.2005.847351

Sidong, L., Yang, S., Weidong, C., Sonia, P., Ron, K., Xiaogang, W., & Dagan, F. (2013). Multifold Bayesian Kernelization in Alzheimer's Diagnosis. Med Image Comput Comput Assist Interv., 16(2), 303–310.

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., & Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, *4*(38), 1–130. doi:10.3310/hta4380 PMID:11134920

Van Calster, B., Timmerman, D., Lu, C., Suykens, J., Valentine, L., Van Holsbeke, C., Amant, F., Vergote, I., & Huffel, S. (2007). Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods. *Ultrasound in Obstetrics & Gynecology*, *29*(5), 496–504. doi:10.1002/uog.3996 PMID:17444557

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer. doi:10.1007/978-1-4757-2440-0

Wahba, G. (1990). Spline Models for Observational Data. SIAM.

Yue, Y. R., Lindquist, M. A., & Loh, J. M. (2012). Meta-Analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *The Annals of Applied Statistics*, *6*(2), 697–718. doi:10.1214/11-AOAS523

*Arti Saxena has more than 14 years of research and teaching experience with specialization in Mathematical Modelling, Fixed Point Theory and Machine Learning.*

*Vijay Kumar has more than 20 years of research and teaching experience with specialization in Machine Learning Intuitionistic fuzzy based models and Engineering Mathematics.*