We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

**4,800**
Open access books available

**122,000**
International authors and editors

**135M**
Downloads

**154**
Countries delivered to

Our authors are among the

**TOP 1%**
most cited scientists

**12.2%**
Contributors from top 500 universities

CLARIVATE ANALYTICS
**BOOK CITATION INDEX**
INDEXED

**WEB OF SCIENCE**™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

# Quantitative Structure-Activity Relationship Modeling and Bayesian Networks: Optimality of Naive Bayes Model

Oleg Kupervasser

Additional information is available at the end of the chapter

**Abstract**

Previously, computational drag design was usually based on simplified laws of molecular physics, used for calculation of ligand's interaction with an active site of a protein-enzyme. However, currently, this interaction is widely estimated using some statistical properties of known ligand-protein complex properties. Such statistical properties are described by quantitative structure-activity relationships (QSAR). Bayesian networks can help us to evaluate stability of a ligand-protein complex using found statistics. Moreover, we are possible to prove optimality of Naive Bayes model that makes these evaluations simple and easy for practical realization. We prove here optimality of Naive Bayes model using as an illustration ligand-protein interaction.

**Keywords:** quantitative structure-activity relationship, Naive Bayes model, optimality, Bayes classifier, Bayesian networks, protein-ligand complex, computational drag design, molecular recognition and binding, ligand-active site of protein, likelihood, probability

## 1. Introduction

The determination within the chapter is based on a paper [1]. Bayes classifiers are broadly utilized right now for recognition, identification, and knowledge discovery. The fields of application are, for case, image processing, personalized medicine [2], chemistry (QSAR (quantitative structure-activity relationship) [3, 4]; see **Figure 1**). The especial importance Bayes Classifiers have in Medical Diagnostics and Bioinformatics. Cogent illustrations of this can be found in the work of Raymer and colleagues [5].
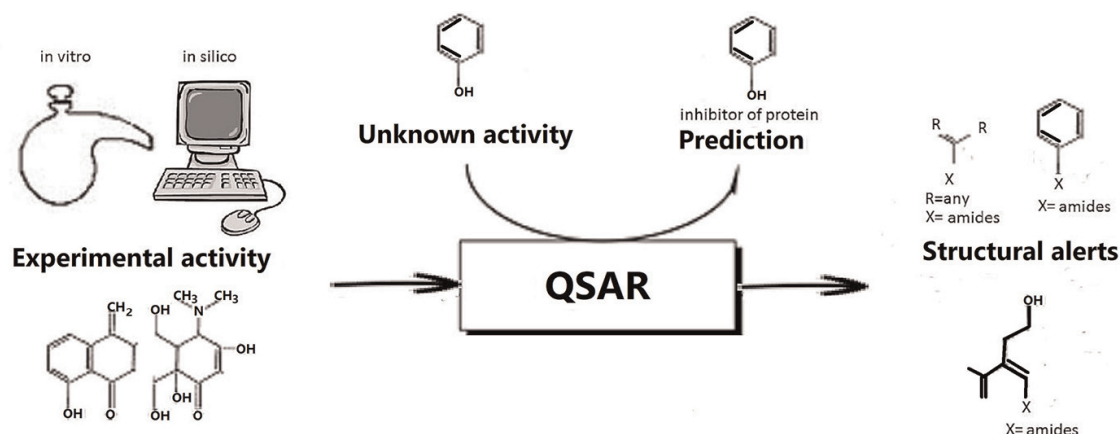
**Figure 1.** Quantitative structure-activity relationship.

Let us give some example of using QSAR from papers [3, 4]:

"Molecular recognition and binding performed by proteins are the background of all biochemical processes in a living cell. In particular, the usual mechanism of drug function is effective binding and inhibition of activity of a target protein. Direct modeling of molecular interactions in protein-inhibitor complexes is the basis of modern computational drug design but is an extremely complicated problem. In the current paradigm, site similarity is recognized by the existence of chemically and spatially analogous regions from binding sites. We present a novel notion of binding site local similarity based on the analysis of complete protein environments of ligand fragments. Comparison of a query protein binding site (target) against the 3D structure of another protein (analog) in complex with a ligand enables ligand fragments from the analog complex to be transferred to positions in the target site, so that the complete protein environments of the fragment and its image are similar. The revealed environments are similarity regions and the fragments transferred to the target site are considered as binding patterns. The set of such binding patterns derived from a database of analog complexes forms a cloudlike structure (fragment cloud), which is a powerful tool for computational drug design."

However, these Bayes classifiers have momentous property—by strange way the Naive Bayes classifier more often than not gives a decent and great description of recognition. More complex models of Bayes classifier cannot progress it significantly [1]. In the paper [6] creators clarify this exceptional property. In any case, they utilize a few suspicions (zero–one misfortune) which diminish all-inclusiveness and simplification of this proof. We allow in this chapter a common verification of Naive Bayes classifier optimality. The induction within the current chapter is comparative to [1]. The consequent attractive consideration of Naive Bayes classifier optimality problem was made in [7, 8]. Be that as it may, shockingly these papers do not incorporate any investigation of the past one [1].

We would like to prove Naive Bayes classifier optimality using QSAR terminology. Indeed, we use QSAR only for clearness; the proof is correct for any field of use of Naive Bayes classifier.

Let us define the essential issue that we attempt to unravel within the chapter. Assume that we have a set of states for a complex of ligand-active site of protein and a set of factors that characterize these states. For each state, we know the likelihood dispersion for each factor. In any case, we have no data of the approximate relationships of the factors. Presently, assume that we know factor values for some test of the state. What is the probability that this test corresponds to some state? It could be a commonplace issue of recognition over a condition of incomplete data.

In the simplest case, we can define two states for "ligand-active site of protein" complex. It is 0 (ligand is not bound to active site of protein) or 1 (ligand is not bound to active site of protein).

The next step is definition of factors (reliabilities below) that characterize strength of a bond for "ligand-active site of protein" complex. Let us grant an illustration of factors (reliabilities below) from experience of QSAR in papers [3, 4]:

"First, consider the protein 5 A°-environment $A = \{a_1, a_2, \ldots a_N\}$ of one ligand atom X in the analog protein, that is, all atoms from the binding site that are in the 5 A°-neighborhood of X. Suppose that the complete target binding site T consists of N′ atoms: $T = \{t_1, t_2, \ldots t_{N'}\}$ and there exists a subset $T_0 \subseteq T$ of size n ($N' \geq n \geq 4$) such that n atoms from $T_0$ are similar to n atoms $A_0 = \{a_{i1}, a_{i2}, \ldots a_{in}\} \subseteq A$ in their chemical types and spatial arrangement. The search for $A_0$ and $T_0$ is performed using a standard clique detection technique in the graph whose nodes represent pairs ($a_i$, $t_i$) of chemically equivalent atoms and edges reflect similarity of corresponding pairwise distances. If the search is successful, the optimal rigid motion superimposing matched protein atoms is applied both to the initial ligand atom X and its complete environment A (**Figure 2(a)** in [3]). The atoms are thus transferred to the target binding site. Then we extend the matching between $A_0$ and $T_0$ by such atom pairs ($a_i$,$t_i$) that $a_i$ and $t_i$ have the same chemical atom type in the coarser 10-type typification mentioned above, and the distance between $t_i$ and the image $a'_i$ of atom $a_i$ is below a threshold. Next, a reliability value R, with $0 \leq R \leq 1$, is assigned to the image X′ of X in the target site and reflects the similarity between the environments of X and its image X′. If the environments are highly similar ($R \approx 1$) we expect that the position of X′ is the place where an atom with chemical type identical to X can be bound by the target, since the environment of X′ contains only atoms required for binding with no "alien" atoms. However, as illustrated in **Figure 2(a)** in [3], the analog site may contain extra binding atoms (shown on the lower side) that decrease the reliability value. In a simple form, the reliability R can be defined as the sum of the number of matched atoms divided by the total number of analog and target atoms in the 5 A°-environments of X and X′, respectively (**Figure 2(b)** in [3]):

$R = 2n/(N + N')$, using the notation presented above. In fact, we use a somewhat more complicated definition that accounts for the quality of spatial superposition of matched atoms and their distance from X′."

We do not want to discuss here these definitions for these factors and states. Our purpose is not the demonstration of effectiveness of these definitions or effectiveness of QSAR. The interested reader can learn it from papers [3, 4] and references inside of these papers. As we said above,
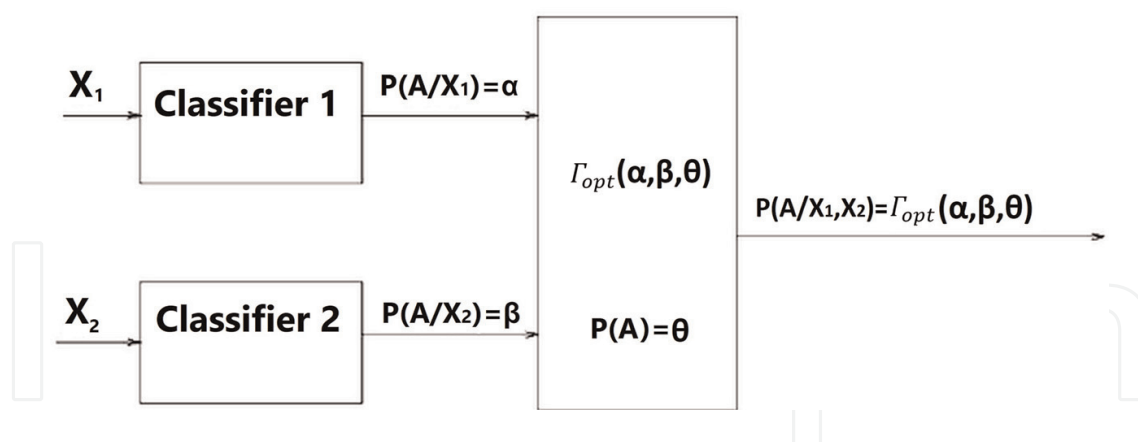
**Figure 2.** Function $\Gamma_{opt}(\alpha, \beta, \theta):[0,1]^3 \rightarrow [0,1]$.

we use QSAR only for clearness; the proof is correct for any field of use of Naive Bayes classifier.

Let us consider the case when no relationships exist between reliabilities. In this case, the Naive Bayes model is a correct arrangement of the issue. We demonstrate in this chapter that for the case that we don't know relationships between reliabilities even approximately—the Naive Bayes model is not correct, but ideal arrangement in a few senses. More point by point, we demonstrate that the Naive Bayes model gives minimal mean error over all conceivable models of relationship. We assume in this confirmation that all relationship models have the same likelihood. We think that this result can clarify the depicted over secretive optimality of Naive Bayes model.

The Chapter is built as described in the following statements. We grant correct numerical description of the issue for two states and two reliabilities in Section 2. We characterize our notations in Section 3. We define general form of conditional likelihood for all conceivable relationships of our reliabilities in Section 4. We characterize the limitations of the functions depicting the relationships in Section 5. We find the formula for an interval between two models of probability (correlation) in Section 6. We discover constraints for our fundamental functions in Section 7. We illuminate our primary issue; we demonstrate Naive Bayes model's optimality for uniform distribution of all conceivable relationships in Section 8. We discover mean error between the Naive Bayes model and a genuine model for uniform distribution of all conceivable relationships in Section 9. We consider the case of more than two states and reliabilities in Section 10. We make conclusions in Section 11.

## 2. Definition of the task

Suppose that A is a state for "ligand-active site of protein" complex. It is 0 (ligand is not bound to active site of protein) or 1 (ligand is not bound to active site of protein). Accept that the *apriori* likelihood $P(A) = P(A = 1)$ is known, and indicate it by $\theta$. Let $X_1, X_2$ be two reliability values (defined above), with values in a set $[0;1]$. However, for generality, we will define $X_1, X_2$ in a set $[-\infty;+\infty]$, but probability density to find $X_1, X_2$ in $[-\infty; 0]$ or $[1;+\infty]$ is equal to zero. We

have the taking after data: $X_1 = x_1$ and $X_2 = x_2$ (gotten through estimation). Moreover, we have two functions, "classifiers," which for given $x_1$ and $x_2$ give us

$$P(A = 1/X_1 = x_1) = P(A/x_1) \equiv \alpha,$$

$$P(A = 1/X_2 = x_2) = P(A/x_2) \equiv \beta.$$

We want to find the likelihood

$$P(A = 1/X_1 = x_1, X_2 = x_2) = P(A/x_1, x_2)$$

in terms of $\alpha, \beta$ and $\theta$. More particularly we wish to discover a function $\Gamma_{opt}(\alpha, \beta, \theta)$ which on the average is the most excellent estimation for $P(A/x_1, x_2)$ in a sense to be characterized expressly within the following consideration (see **Figure 2**).

## 3. Notation and preliminaries

$\rho_{X_1, X_2}(x_1, x_2)$—joint PDF (probability density function) for $X_1$ and $X_2$.

$\rho_{X_1, X_2/A}(x_1, x_2) = h(x_1, x_2)$—joint PDF for $X_1$ and $X_2$, known $A = 1$. In terms of $h(x_1, x_2)$ and $\theta$, we can find $P(A/x_1, x_2)$ as follows:

$$P(A/x_1, x_2) = \frac{\theta h(x_1, x_2)}{\theta h(x_1, x_2) + (1 - \theta)\overline{h}(x_1, x_2)}, \tag{1}$$

here

$\overline{h}(x_1, x_2) \equiv \rho_{X_1, X_2/\overline{A}}(x_1, x_2)$—joint PDF for $X_1$ and $X_2$, known $A = 0$.

We can find

$$\rho_{X_1}(x_1) = \int_{-\infty}^{+\infty} \rho_{X_1, X_2}(x_1, x_2) dx_2,$$

$$\rho_{X_2}(x_2) = \int_{-\infty}^{+\infty} \rho_{X_1, X_2}(x_1, x_2) dx_1,$$

$$h_1(x_1) \equiv \rho_{X_1/A}(x_1) = \int_{-\infty}^{+\infty} h(x_1, x_2) dx_2,$$

$$h_2(x_2) \equiv \rho_{X_2/A}(x_2) = \int_{-\infty}^{+\infty} h(x_1, x_2) dx_1,$$

$$\overline{h}_1(x_1) \equiv \rho_{X_1/\overline{A}}(x_1) = \int_{-\infty}^{+\infty} \overline{h}(x_1, x_2) dx_2,$$

$$\overline{h}_2(x_2) \equiv \rho_{X_2/\overline{A}}(x_2) = \int_{-\infty}^{+\infty} \overline{h}(x_1, x_2) dx_1.$$

## 4. Generic form of $P(A/x_1, x_2)$

Let us define the function $g(x_1, x_2)$ and $\overline{g}(x_1, x_2)$

$$g(x_1, x_2) \equiv \frac{h(x_1, x_2)}{h_1(x_1)h_2(x_2)},$$

$$\overline{g}(x_1, x_2) \equiv \frac{\overline{h}(x_1, x_2)}{\overline{h}_1(x_1)\overline{h}_2(x_2)}.$$

Let us say that if $X_1$ and $X_2$ are conditionally independent, i.e.,

$$h(x_1, x_2) = \rho_{X_1 X_2/A}(x_1, x_2) = \rho_{X_1/A}(x_1)\rho_{X_2/A}(x_2)$$
$$= h_1(x_1)h_2(x_2),$$

then

$$g(x_1, x_2) = \overline{g}(x_1, x_2) = 1.$$

Let us define the following *monotonously nondecreasing* probability distribution functions:

$$H_1(x_1) \equiv \int_{-\infty}^{x_1} h_1(z) dz,$$

$$H_2(x_2) \equiv \int_{-\infty}^{x_2} h_2(z) dz,$$

$$\overline{H}_1(x_1) \equiv \int_{-\infty}^{x_1} \overline{h}_1(z) dz,$$

$$\overline{H}_2(x_2) \equiv \int_{-\infty}^{x_2} \overline{h}_2(z) dz.$$

Take attention that since $H_1(x_1), H_2(x_2), \overline{H}_1(x_1)$ and $\overline{H}_2(x_2)$ are monotonous (At this point we can assume that $h_1(x_1), h_2(x_2), \overline{h}_1(x_1), \overline{h}_2(x_2) > 0$ so that $H_1(x_1), H_2(x_2), \overline{H}_1(x_1)$ and $\overline{H}_2(x_2)$ are monotonously increasing. However, such limitation will be unnecessary as we will see within the following conclusion.), the inverse functions $H_1^{-1}(x_1), H_2^{-1}(x_2), \overline{H}_1^{-1}(x_1)$ and $\overline{H}_2^{-1}(x_2)$ must exist. As a result, we can characterize

$$J(a,b) \equiv g\left(H_1^{-1}(a), H_2^{-1}(b)\right),$$

$$\overline{J}(a,b) \equiv \overline{g}\left(\overline{H}_1^{-1}(a), \overline{H}_2^{-1}(b)\right).$$

To be brief, let us use the following concise designation:

$$J \equiv J(H_1(x_1), H_2(x_2)) = g\left(H_1^{-1}(H_1(x_1))\right), H_2^{-1}(H_2(x_2)) = g(x_1, x_2),$$

$$\overline{J} \equiv \overline{J}(\overline{H}_1(x_1), \overline{H}_2(x_2)) = \overline{g}\left(\overline{H}_1^{-1}(\overline{H}_1(x_1)), \overline{H}_2^{-1}(\overline{H}_2(x_2))\right) = \overline{g}(x_1, x_2).$$

By the definition

$$h(x_1, x_2) = J h_1(x_1) h_2(x_2), \tag{2}$$

$$\overline{h}(x_1, x_2) = \overline{J}\, \overline{h}_1(x_1) \overline{h}_2(x_2). \tag{3}$$

We currently obtain

$$h_1(x_1) \equiv \rho_{X_1/A}(x_1) = \frac{\rho_{X_1}(x_1) P(A/x_1)}{P(A)} = \frac{\alpha \rho_{X_1}(x_1)}{\theta}, \tag{4}$$

$$h_2(x_2) \equiv \rho_{X_2/A}(x_2) = \frac{\rho_{X_2}(x_2) P(A/x_2)}{P(A)} = \frac{\beta \rho_{X_2}(x_2)}{\theta}, \tag{5}$$

$$\overline{h}_1(x_1) \equiv \rho_{X_1/\overline{A}}(x_1) = \frac{\rho_{X_1}(x_1) P(\overline{A}/x_1)}{P(\overline{A})} = \frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta}, \tag{6}$$

$$\overline{h}_2(x_2) \equiv \rho_{X_2/\overline{A}}(x_2) = \frac{\rho_{X_2}(x_2) P(\overline{A}/x_2)}{P(\overline{A})} = \frac{(1-\alpha)\rho_{X_2}(x_2)}{1-\theta}. \tag{7}$$

As a result from Eqs. (2) and (3)

$$h(x_1, x_2) = J \frac{\alpha \beta \rho_{X_1}(x_1) \rho_{X_2}(x_2)}{\theta^2},$$

$$\overline{h}(x_1, x_2) = \overline{J} \frac{(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{(1-\theta)^2}.$$

Now from Eq. (1)

$$P(A/x_1, x_2) = \frac{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2) + \frac{\bar{J}}{(1-\theta)}(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)}$$

(8)

$$= \frac{\alpha\beta}{\alpha\beta + \frac{\bar{J}}{J}\frac{\theta}{1-\theta}(1-\alpha)(1-\beta)}.$$

Note, that for values of $J = \bar{J} = 1$ (conditional independence of $x_1$ and $x_2$) equation (8) becomes the exact solution for the optimal model:

$$\Gamma(\alpha, \beta, \theta) = P(A/x_1, x_2).$$

## 5. Limitations for the functions $J(a, b)$ and $\bar{J}(a, b)$

We can write

$$h_1(x_1) = \int_{-\infty}^{+\infty} J(H_1(x_1), H_2(x_2))h_1(x_1)h_2(x_2)dx_2.$$

(9)

As a result

$$1 = \int_{-\infty}^{+\infty} J(H_1(x_1), H_2(x_2))h_2(x_2)dx_2 = \int_0^1 J(H_1(x_1), H_2(x_2))dH_2(x_2).$$

(10)

Thus, we obtain the following condition:

$$\int_0^1 J(a, b)db = 1,$$

(11)

and similarly

$$\int_0^1 J(a, b)da = 1.$$

(12)

Similarly, we can get

$$\int_0^1 \bar{J}(a,b)da = 1$$

$$\int_0^1 \bar{J}(a,b)db = 1. \tag{13}$$

Obviously

$$J(a,b), \bar{J}(a,b) \geq 0, \tag{14}$$

$$\int_0^1\int_0^1 J(a,b)dadb = \int_0^1\int_0^1 \bar{J}(a,b)dadb = 1. \tag{15}$$

All the solutions of Eqs. (11)–(15) together with (8) can define the set of all possible realizations of $P(A/x_1, x_2)$.

Let us give some example of a solution of (11), (12) and (14), (15):

Let $\rho(x)$ be a function such that $\rho(x) \geq 0$ and $\int_0^1 \rho(x)dx = 1$

$$J(a,b) = \begin{cases} \rho(a-b) & , a \geq b \\ \rho(a-b+1) & , a < b' \end{cases}$$

satisfy Eqs. (11), (12), (14), and (15).

# 6. Definition of distance

We define the distance between the proposed approximation of $P(A/x_1, x_2)$-$\Gamma(\alpha, \beta, \theta)$ and the actual function $P(A/x_1, x_2)$ as follows:

$$\|\Gamma(\alpha, \beta, \theta) - P(A/x_1, x_2)\| = \int_{-\infty}^{+\infty}\int \rho_{X_1 X_2}(x_1, x_2)[\Gamma(\alpha, \beta, \theta) - P(A/x_1, x_2)]^2 dx_1 dx_2.$$

Now we have from Eqs. (2) and (3) and Eqs. (4)–(7)

$$\rho_{X_1 X_2}(x_1, x_2) = \theta h(x_1, x_2) + (1-\theta)\bar{h}(x_1, x_2)$$

$$= \theta J h_1(x_1)h_2(x_2) + (1-\theta)\bar{J}\bar{h}_1(x_1)\bar{h}_2(x_2)$$

$$= \left[J\frac{\alpha\beta}{\theta} + \bar{J}\frac{(1-\alpha)(1-\beta)}{(1-\theta)}\right]\rho_{X_1}(x_1)\rho_{X_2}(x_2),$$

$$\|\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)\| = \int\limits_{-\infty}^{+\infty}\int \rho_{X_1}(x_1)\rho_{X_2}(x_2)\left[J\frac{\alpha\beta}{\theta} + \bar{J}\frac{(1-\alpha)(1-\beta)}{(1-\theta)}\right] \times (\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2))^2 dx_1 dx_2$$

$$= \int\limits_{0}^{1}\int\limits_{0}^{1}\left[\frac{J\alpha\beta}{\theta} + \frac{\bar{J}(1-\alpha)(1-\beta)}{(1-\theta)}\right] \times (\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2))^2 dF_1(x_1)dF_2(x_2).$$

$$(16)$$

Here

$$F_1(x_1) = \int_{-\infty}^{x_1} \rho_{X_1}(z)dz,$$

$$F_2(x_2) = \int_{-\infty}^{x_2} \rho_{X_2}(z)dz.$$

## 7. Constraints for basic functions

We will consider further all functions with arguments $1 \geq F_1, F_2 \geq 0$, but not $x_1, x_2$. We have six functions of $F_1, F_2$, which define Eq. (16): $J, \bar{J}, H_1, H_2, \alpha, \beta$. Let us to write the functions by help these functions ($F_1, F_2$) and find restrictions for these functions:

$$\alpha = P(A/x_1) = \theta h_1(x_1)/\rho_{X_1}(x_1) = \theta(dH_1/dx_1)/(dF_1/dx_1) = \theta\frac{dH_1}{dF_1}.$$

By the same way

$$\beta = \theta\frac{dH_2}{dF_2}.$$

We know that functions $H_1, F_1, H_2, F_2$ are cumulative distribution functions of $x_1, x_2$, correspondingly. These functions are *monotonously nondecreasing* and change from 0 to 1 from the definition of cumulative distribution functions. Therefore, we can conclude the following restraints for functions $H_1, H_2$ as functions of $F_1, F_2$ exist:

$$H_1(1) = H_2(1) = 1,$$

$$H_1(0) = H_2(0) = 0,$$

$$0 \leq \alpha = \theta\frac{dH_1}{dF_1}, \beta = \theta\frac{dH_2}{dF_2} \leq 1,$$

$$0 \leq \theta \leq 1,$$

$$\overline{H}_1(x_1) = \int_{-\infty}^{x_1} \overline{h}_1(x_1)dx_1$$

$$= \int_{-\infty}^{x_1} \frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta} dx_1$$

$$= \frac{1}{1-\theta}\int_{-\infty}^{x_1} \rho_{X_1}(x_1)dx_1 - \frac{\theta}{1-\theta}\int_{-\infty}^{x_1} \frac{\alpha\rho_{X_1}(x_1)}{\theta} dx_1$$

$$= \frac{F_1}{1-\theta} - \frac{\theta}{1-\theta}H_1(x_1).$$

By the same way

$$\overline{H}_2(x_2) = \frac{F_2}{1-\theta} - \frac{\theta}{1-\theta}H_2(x_2),$$

$$J(H_1(F_1), H_2(F_2)) : J(H_1(F_1), H_2(F_2)) \geq 0$$

$$\int_0^1 J(a,b)db = 1$$

$$\int_0^1 J(a,b)da = 1,$$

$$\overline{J}(\overline{H}_1(F_1), \overline{H}_2(F_2)) : \overline{J}(\overline{H}_1(F_1), \overline{H}_2(F_2)) \geq 0$$

$$\int_0^1 \overline{J}(a,b)db = 1$$

$$\int_0^1 \overline{J}(a,b)da = 1,$$

$$P(A/x_1, x_2) = \frac{\frac{J\alpha\beta}{\theta}}{\frac{J\alpha\beta}{\theta} + \frac{\overline{J}(1-\alpha)(1-\beta)}{1-\theta}}. \tag{17}$$

## 8. Optimization

We shall find the best approximation of $\Gamma(\alpha, \beta, \theta)$ as follows:

$$min_{\Gamma(\alpha,\beta,\theta)} E\big[\|\Gamma(\alpha, \beta, \theta) - P(A/x_1, x_2)\|\big] \to \Gamma(\alpha, \beta, \theta),$$

where the expected value (or expectation or mathematical expectation or mean or the first moment) $E[\ldots]$ is taken with respect to the joint PDF of possible realizations of $J, \overline{J}, \alpha, \beta, H_1, H_2$ for given $F_1$ and $F_2$.

For the sake of brevity, we denote

$$C \equiv \frac{J\alpha\beta}{\theta} + \frac{\overline{J}(1-\alpha)(1-\beta)}{1-\theta}$$

$$D \equiv \frac{J\alpha\beta}{\theta}.$$

Then from Eqs. (17) and (16)

$$\|\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)\| = \int_0^1\int_0^1 C(\Gamma(\alpha,\beta,\theta) - D/C)^2 dF_1 dF_2$$

$$= \int_0^1\int_0^1 dF_1 dF_2 [D^2 C + \Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D].$$

(18)

Thus

$$min_{\Gamma(\alpha,\beta,\theta)} E[\|\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)\|]$$

$$= min_{\Gamma(\alpha,\beta,\theta)} E\left[\int_0^1\int_0^1 dF_1 dF_2 [D^2 C + \Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]\right]$$

$$= min_{\Gamma(\alpha,\beta,\theta)} E\left[\int_0^1\int_0^1 dF_1 dF_2 [D^2 C]\right]$$

$$+ min_{\Gamma(\alpha,\beta,\theta)} E\left[\int_0^1\int_0^1 dF_1 dF_2 [\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]\right]$$

$$= Const + min_{\Gamma(\alpha,\beta,\theta)} E\left[\int_0^1\int_0^1 dF_1 dF_2 [\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D]\right].$$

(19)

It remains to calculate the expected value in Eq. (19).

We have by obvious assumptions

$$\rho_{J,\bar{J},\alpha,\beta,H_1 H_2/F_1,F_2}(J,\bar{J},\alpha,\beta,H_1,H_2/F_1,F_2) = \rho_{J/H_1,H_2}(J/H_1,H_2)\rho_{\bar{J}/\overline{H}_1,\overline{H}_2}(\bar{J}/\overline{H}_1,\overline{H}_2)\rho_{\alpha/F_1}(\alpha/F_1)$$

$$\times \rho_{H_1/\alpha F_1}(H_1/\alpha,F_1)\rho_{\beta/F_2}(\beta/F_2)\rho_{H_2/\beta,F_2}(H_2/\beta,F_2).$$

(20)

Lemma 1

$$E[J(a,b)] = \int_0^{+\infty} \rho_{J(a,b)/a,b}(J(a,b)/a,b)J(a,b)dJ = 1,$$

$$E[\bar{J}(a,b)] = \int_0^{+\infty} \rho_{\bar{J}(a,b)/a,b}(\bar{J}(a,b)/a,b)\bar{J}(a,b)d\bar{J} = 1.$$

Proof: We can take into the consideration the function $\rho_{J(a,b)/a,b}$. The domain of the function $J(a,b)$ is square $0 \le a, b \le 1$. By dividing this square into small squares $(i,j)$, we can get sampling of the function $J$. Then, on every square $i, j$, we can define the value of the function $J_{ij}$. We can write the following restraints for function $J(***)$:

$$J_{ij} \geq 0,$$

$$\frac{1}{N} \sum_{i=1}^{N} J_{ij} = 1,$$

$$\frac{1}{N} \sum_{j=1}^{N} J_{ij} = 1.$$

Here $i = 1, ..., N, j = 1, ..., N$.

All matrixes $(J_{ij})$ that satisfy the above conditions have the same probability. So we can define probability density function

$$\rho\left(J_{11}, ..., J_{ij}, ..., J_{NN}\right).$$

This density function should be symmetric according to transpositions of columns and rows of the matrix $(J_{ij})$, because the density function has the same probability for all matrixes $(J_{ij})$ that satisfy the above conditions. Indeed, these conditions are also symmetric according to transpositions of columns and rows of matrix $(J_{ij})$. From symmetry conditions that define this function $(\rho)$ according to transpositions of columns and rows of matrix $(J_{ij})$, it is possible to conclude that this function $(\rho)$ also does not transform according to these transpositions.

We can consider function $\rho_{u/ij}(u/ij)$ that is a discretization of the function $\rho_{J(a,b)/a,b}(J(a,b)/a,b)$:

$$\rho_{u/ij}(u/ij) = \int ... \int_0^{+\infty} \rho\left(J_{11}, ...J_{nk}, ...J_{ij} = u, ..., J_{NN}\right) \prod_{(lm) \neq (ij)} dJ_{lm}.$$

We can transpose columns and rows $(J_{ij})$ in such a way that element $J_{ij}$ will be replaced by the other element $J_{nk}$, and after it the function $\rho(J_{11}, ...)$ will not be transformed. So from the above equation, we can get

$$\rho_{u/ij}(u/ij) = \int ... \int_0^{+\infty} \rho\left(J_{11}, ...J_{nk}, ...J_{ij} = u, ..., J_{NN}\right) \prod_{(lm) \neq (nk)} dJ_{lm} = \rho_{u/nk}(u/nk).$$

From this equation we can conclude that $\rho_{u/ij}(u/ij)$ does not depend on $ij$ so $\rho_{J/a,b}(J/a,b)$ does not depend on $ab$ and

$$\rho_{J/a,b}(J/a,b) = \rho_J(J),$$

and

$$E[J(a,b)] = \int_0^{+\infty} \rho_J(J)J dJ = Const, \qquad (21)$$

From

$$\int_0^1 \int_0^1 J(a,b)\,da\,db = 1,$$

we can conclude that

$$\int_0^1 \int_0^1 E[J(a,b)]\,da\,db = 1.$$

So we can obtain that $Const = 1$ in Eq. (21).

Lemma 2: Probability distribution functions $\alpha$ and $\beta$ do not depend on $F_1$ and $F_2$:

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha),$$

$$\rho_{\beta/F_2}(\beta/F_2) = \rho_\beta(\beta).$$

Proof: Let us make sampling of the function $\alpha(F_1)$ by dividing the domain of this function $F_1$, $[0,1]$ on intervals of $1/N$, $N \gg 1$. Then restriction conditions for $\alpha_k$, $k = 1, \ldots, N$

$$0 \le \alpha_k \le 1,$$

$$\frac{1}{N} \sum_{k=1}^N \alpha_k = \int_0^1 \theta\, dH_1(F_1)\, dF_1\, dF_1 = \theta.$$

All columns $(\alpha_k)$ that are satisfied by these conditions have equal probability. We can consider respective function $\rho(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_l, \ldots, \alpha_N)$. From symmetry conditions that define this function according to transpositions $\alpha_k \to \alpha_l$, function $\rho(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_l, \ldots, \alpha_N)$ also does not transform according to such transpositions. As a result, it is possible to write

$$\rho_k(u) = \int_0^1 \rho(\alpha_1, \ldots, \alpha_k = u, \ldots, \alpha_l, \ldots, \alpha_N) \prod_{n \neq k} d\alpha_n$$

$$= \int_0^1 \rho(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_l = u, \ldots, \alpha_N) \prod_{n \neq l} d\alpha_n$$

$$= \rho_l(u).$$

From this equation, we can conclude that function $\rho_{\alpha/F_1}(\alpha/F_1)$ does not depend on $F_1$:

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha).$$

From (20) we obtain

$$E\left[\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D\right] = \int_0^1\int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta$$

$$\times \int_0^1\int_0^1 \rho_{H_1/\alpha,F_1}(H_1/\alpha,F_1)\rho_{H_2/\beta,F_2}(H_2/\beta,F_2)dH_1 dH_2 \int_0^\infty\int_0^\infty \rho_J(J)\rho_{\bar{J}}(\bar{J})$$

$$\times \left[\Gamma^2(\alpha,\beta,\theta)\left[\frac{J\alpha\beta}{\theta} + \frac{\bar{J}(1-\alpha)(1-\beta)}{1-\theta}\right] - 2\Gamma(\alpha,\beta,\theta)\frac{J\alpha\beta}{\theta}\right]dJd\bar{J}$$

$$= \int_0^1\int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta$$

$$\times \left[\Gamma^2(\alpha,\beta,\theta)\left[\frac{E[J]\alpha\beta}{\theta} + \frac{E[\bar{J}](1-\alpha)(1-\beta)}{1-\theta}\right] - 2\Gamma(\alpha,\beta,\theta)\frac{E[J]\alpha\beta}{\theta}\right].$$

Let us define

$$\overline{C} = \frac{\alpha\beta}{\theta} + \frac{(1-\alpha)(1-\beta)}{1-\theta},$$

$$\overline{D} = \frac{\alpha\beta}{\theta}.$$

By Lemma 1, $E[J] = E[\bar{J}] = 1$. Hence

$$E\left[\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D\right] = \int_0^1\int_0^1 \left[\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}\right]\rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta.$$

It remains to find

$$min_{\Gamma(\alpha,\beta,\theta)} \int_0^1\int_0^1 dF_1 dF_2 \int_0^1\int_0^1 d\alpha d\beta \rho_\alpha(\alpha)\rho_\beta(\beta)\left[\Gamma^2(\alpha,\beta,\theta))\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}\right]. \qquad (22)$$

Since

$$\rho_\alpha(\alpha)\rho_\beta(\beta) \geq 0,$$

if the expression in square brackets is minimized at each point, then the whole integral in Eq. (22) is minimized. Thus, we may proceed as follows:

$$\frac{\partial}{\partial \Gamma}\left[\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}\right] = 2\Gamma(\alpha,\beta,\theta)\overline{C} - 2\overline{D} = 0.$$

Hence the optimum $\Gamma(\alpha,\beta,\theta)$ is given by

$$\Gamma_{opt}(\alpha,\beta,\theta) = \frac{\overline{D}}{\overline{C}} = \frac{\frac{\alpha\beta}{\theta}}{\frac{\alpha\beta}{\theta} + \frac{(1-\alpha)(1-\beta)}{1-\theta}}.$$

# 9. Mean distance between the proposed approximation of $P(A/x_1,x_2) - \Gamma(\alpha,\beta,\theta)$ and the actual function $P(A/x_1,x_2)$

The mean distance from (18) is

$$DIS = E\left[\left\|\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)\right\|\right]$$

$$= \int_0^1 \int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta\left[\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}\right] + Const,$$

where *Const* in this equation is defined by

$$Const = E\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho_{X_1X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1 dx_2\right].$$

From this equation we can find boundaries of the *Const*. From $0 \le P(A/x_1,x_2) \le 1$ we can conclude

$$Const \le E\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho_{X_1,X_2,}(x_1,x_2)P(A/x_1,x_2)dx_1 dx_2\right] = E[\theta] = \theta.$$

The second condition is

$$0 \le E\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2) - \theta]^2 dx_1 dx_2\right]$$

$$= E\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho_{X_1,X_2}(x_1,x_2)\left[P(A/x_1,x_2)^2 + \theta^2 - 2P(A/x_1,x_2)\theta\right]dx_1 dx_2\right]$$

$$= E\left[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1 dx_2\right] - \theta^2.$$

So from these two equations, we can conclude

$$\theta^2 \leq Const \leq \theta.$$

In the next step, we would like find function $\rho_\alpha(\alpha)$ $(\rho_\beta(\beta))$ in the equation for *DIS*.

Restrictions for function $\alpha(F_1), 0 \leq F_1 \leq 1$ are the following:

$$\int_0^1 \alpha(F_1)dF_1 = \theta,$$

$$0 \leq \alpha(F_1) \leq 1.$$

In discrete form (for $N \to \infty$), we can rewrite $\alpha_{set} = \{\alpha_1, \alpha_2, ..., \alpha_N\}$

$$\frac{1}{N}\sum_{i=1}^N \alpha_i = \theta,$$

$$0 \leq \alpha_i \leq 1, i = 1, 2, ..., N.$$

Let us define a function $U(\alpha_{set})$ in the following way:

$$U(\alpha_{set}) = \begin{cases} \sum_{i=1}^N \alpha_i & \text{for } 0 \leq \alpha_i \leq 1 \quad , i = 1, 2, ..., N \\ +\infty & \text{otherwise} \end{cases},$$

$$U(\alpha_{set}) = \sum_{i=1}^N U_i(\alpha_i),$$

$$U_i(\alpha_i) = \begin{cases} \alpha_i & \text{for } 0 \leq \alpha_i \leq 1 \\ +\infty & \text{otherwise} \end{cases}.$$

Then the function that satisfies equal probability distribution with considering restrictions (i) and (ii) is the following:

$$\rho_{\alpha_{set}}(\alpha_{set}) = \frac{1}{C}\delta(U(\alpha_{set}) - N\theta). \tag{23}$$

here $\delta$ is the Dirac delta function.

We can define the constant $C$ by

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_{set})d\alpha_1...d\alpha_N = 1.$$

It can be proved for $N \to \infty$ that distribution (23) is equal to the following distribution (from "statistical mechanics" [9]; transform from microcanonical to canonical distribution):

$$\rho_{\alpha_{set}}(\alpha_{set}) = \frac{1}{Z} e^{-KU(\alpha_{set})}.$$

Here we can find $Z$ and $K$ from the following equations:

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_{set}) d\alpha_1 \ldots d\alpha_N = 1, \tag{24}$$

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} U(\alpha_{set}) \rho_{\alpha_{set}}(\alpha_{set}) d\alpha_1 \ldots d\alpha_N = N\theta. \tag{25}$$

Quest function $\rho_\alpha(\alpha)$ can be found by

$$\rho_\alpha(\alpha) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_1, \ldots, \alpha_j = \alpha, \ldots, \alpha_N) \prod_{i=1, i \neq j}^{N} d\alpha_i = \frac{1}{D} e^{-KU_j(\alpha_j = \alpha)}, \tag{26}$$

where

$$D^N = Z. \tag{27}$$

From Eqs. (24) and (25), we can find

$$\frac{1}{Z} = \left( \frac{K}{1 - e^{-K}} \right)^N, \tag{28}$$

$$\theta = \Lambda(K), \tag{29}$$

where $\Lambda(K)$ is the decreasing function

$$\Lambda(K) = \begin{cases} 1 & \text{for } K = -\infty \\ 0 & \text{for } K = +\infty \\ 1/2 & \text{for } K = 0 \\ \dfrac{1}{K} - \dfrac{1}{e^K - 1} & \text{otherwise} \end{cases}.$$

If $K$ is the root of Eq. (29), we can write from Eqs. (26)–(29) for function $\rho_\alpha(\alpha)$

$$\rho_\alpha(\alpha) = \begin{cases} \begin{cases} \text{For} \qquad K = 0 \\ 1 \qquad \text{for } 0 \le \alpha \le 1 \\ 0 \qquad \text{otherwise } \alpha \end{cases} \\ \begin{cases} \text{For} \qquad K = +\infty \\ 2\delta(\alpha) \qquad 0 \le \alpha \le 1 \\ 0 \qquad \text{otherwise } \alpha \end{cases} \\ \begin{cases} \text{For} \qquad K = -\infty \\ 2\delta(\alpha - 1) \qquad 0 \le \alpha \le 1 \\ 0 \qquad \text{otherwise } \alpha \end{cases} \\ \begin{cases} \text{For} \qquad \text{otherwise } K \\ \dfrac{1}{D}e^{-K\alpha} \qquad 0 \le \alpha \le 1 \\ 0 \qquad \text{otherwise } \alpha \end{cases} \end{cases} ,$$

where $2\displaystyle\int_0^1 \delta(\alpha - 1)d\alpha = 2\int_0^1 \delta(\alpha)d\alpha = 1$ and $\frac{1}{D} = \frac{K}{1-e^{-K}}$.

## 10. The case of more than two states $A$ and reliabilities $X$

Let A be a state, with values in set 0, 1, …, L. This number can characterize strength of a bond. Assume that the *apriori* probability $P(A = i)$ is known, and denote it by $\theta_i$; here $i = 1, …, L$. Let $X_1, …, X_K$ be random variables, with values in some set, say $] -\infty; +\infty[$. We have the following information: $X_1 = x_1,…,X_K = x_K$ (obtained through measurement). Furthermore, we have systems, "classifiers," which for given $x_1,…,x_K$ produce

$$P(A = i/X_j = x_j) \equiv \alpha_{ij}.$$

We want to find the probability $P(A = i/X_1 = x_1, …, X_K = x_K)$ in terms of $\alpha_{ij}$ and $\theta_i$. In more detail, we want to find a function $\Gamma_{opt,M}(\alpha_{ij}, \theta_i)$, which is the best approximation for $P(A = M/x_1, …, x_K)$ on the average. By the same way, in case of two variables, it is possible to find that the $\Gamma_{opt,M}(\alpha_{ij}, \theta_i)$ can be defined by the following equation:

$$\Gamma_{opt,M}(\alpha_{ij}, \theta_i) = \frac{\left(\prod_{j=1}^K \alpha_{Mj}\right)/\theta_M^{K-1}}{\sum_{i=1}^L \left(\prod_{j=1}^K \alpha_{ij}\right)/\theta_i^{K-1}}.$$

We have evidential restraints for $\alpha_{ij}, \theta_i$

$$0 \leq \alpha_{ij} \leq 1$$

$$\sum_{i=1}^{L} \alpha_{ij} = 1,$$

$$0 \leq \theta_i \leq 1$$

$$\sum_{i=1}^{L} \theta_i = 1.$$

## 11. Conclusions

Using as an illustration the QSAR, we demonstrated effectively that the Naive Bayes model gives minimal mean error over uniform dispersion of all conceivable relationships between characteristic reliabilities. This result can clarify the portrayed over secretive optimality of Naive Bayes model. We too found the mean error that the Naive Bayes model gives for uniform distribution of all conceivable relationships of reliabilities.

Medicinal chemistry (quantitative structure-activity relationships, QSAR) prediction increasingly relies on Bayesian network-based methods. Its importance derives partly from the difficulty and inaccuracies of present quantum chemical models (e.g., in SYBYL and other software) and from the impracticality of sufficient characterization of structure of drug molecules and receptor active sites, including vicinal waters in and around hydrophobic pockets in active sites. This is particularly so for biologicals (protein and nucleic acid APIs (nucleic acid active pharmaceutical ingredients)) and target applications that exhibit extensive inter-receptor trafficking, genomic polymorphisms, and other system biology phenomena. The effectiveness and accuracy of Bayesian methods for drug development likewise depend on certain prerequisites, such as an adequate distance metric by which to measure similarity/difference between combinatorial library molecules and known successful ligand molecules targeting a particular receptor and addressing a particular clinical indication. In this connection, the distance metric proposed in Section 6 of the chapter manuscript and the associated Lemmas and Proofs are of substantial value in the future of high-throughput screening (HTS) and medicinal chemistry.

However, our purpose here was not demonstration of effectiveness of these definitions or effectiveness of QSAR. The interested reader can learn it from papers [3, 4] and references inside of these papers. As we said above, we use QSAR only for clearness; the proof is correct for any field of use of Naive Bayes classifier.

## Author details

Oleg Kupervasser

Address all correspondence to: olegkup@yahoo.com

Department of Mathematics, Ariel University, Ariel, Israel

# References

[1]  Kupervasser O. The mysterious optimality of Naive Bayes: Estimation of the probability in the system of "classifiers". Pattern Recognition and Image Analysis. 2014;**24**(1):1-10. http://arxiv.org/abs/cs/0202020v1; 2002

[2]  Maslennikov ED, Sulimov AV, Savkin IA, Evdokimova MA, Zateyshchikov DA, Nosikov VV, et al. An intuitive risk factors search algorithm: Usage of the Bayesian network technique in personalized medicine. Journal of Applied Statistics. 2015;**42**(1):71-87

[3]  Ramensky V, Sobol A, Zaitseva N, Rubinov A, Zosimov V. A novel approach to local similarity of protein binding sites substantially improves computational drug design results. Proteins: Structure, Function, and Bioinformatics. 2007;**69**(2):349-357

[4]  Nikitin S, Zaitseva N, Demina O, Solovieva V, Mazin E, Mikhalev S, et al. A very large diversity space of synthetically accessible compounds for use with drug design programs. Journal of Computer-Aided Molecular Design. 2005;**19**:47-63

[5]  Raymer ML, Doom TE, Kuhn LA, Punch WF. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems, Man, and Cybernetics. 2003;**33B**:802

[6]  Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning. 1997;**29**:103

[7]  Zhang H. The optimality of Naive Bayes. In: FLAIRS Conference; 2004. Available from: http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf

[8]  Kuncheva LI. On the optimality of Naive Bayes with dependent binary features. Pattern Recognition Letters. 2006;**27**:830

[9]  Landau LD, Lifshitz EM. Statistical Physics. Vol. 5. United Kingdom: Elsevier Science Technology; 1996