The Mysterious Optimality of Naive Bayes: Estimation of the Probability in the System of "Classifiers"

Oleg Kupervasser
Department of Chemical Physics,The Weizmann Institute of Science,Rehovot 76100, Israel

Bayes Classifiers are widely used currently for recognition, identification and knowledge discovery. The fields of application are, for example, image processing, medicine, chemistry (QSAR). However, by mysterious way the Naive Bayes Classifier usually gives a very nice and good presentation of recognition. More complex models of Bayes Classifier cannot improve it considerably. We demonstrate here a very nice and simple proof of the Naive Bayes Classifier optimality that can explain this interesting fact. The derivation in the current paper is based on a paper of the author written in 2002.

PACS numbers: PACS numbers 47.27.Gs, 47.27.Jv, 05.40.+j

I. INTRODUCTION

The derivation in the current paper is based on a paper of the author written in 2002 [1] . Bayes Classifiers are widely used currently for recognition, identification and knowledge discovery. The fields of application are, for example, image processing, medicine, chemistry (QSAR). The special significance such Classifiers have in Medical Diagnostics and Bioinformatics. Very nice examples can be found in paper [2]. However, these Bayes Classifiers have remarkable property - by mysterious way the Naive Bayes Classifier usually gives a very nice and good presentation of recognition. More complex models of Bayes Classifier [3] cannot improve it considerably.

Let us give some example from practices of author. The first example was recognition of digits written by hand. Every such digit can be characterized by set of variables. The second example is defects on a computer screen, such as scratches, air bubbles, cavities, spots. They can be characterized by set of variables, for example, square of circumscribed ellipse, its eccentricity and so on. The third example is medical diagnostics. We must recognize the diseases on basis of medical symptoms. The all three examples had the same property: in spite of the fact that correlations exist between characteristic variables, the Naive Bayes model gave the excellent result. Moreover, this result could not be improved considerably by using more complex model with some correlations between characteristic variables. Sometimes these correlations (if they are found with errors) can make the model even worse.

In the paper [3] authors explain this remarkable property. However, they use some assumption (Zero-One Loss) which decreases universality and generality of this consideration. We give in this paper a general proof Naive Bayes Classifier optimality. The derivation in the current paper is similar to [1] (2002). The subsequent interesting development of the problem was made in [4] (2004), [5] (2006). However, unfortunately these papers do not include any analysis of previous one [1].

Let us formulate shortly the basic problem that we try to solve in the paper. Suppose that we have a set of some objects and a set of variables that characterize these objects. For every object, we know probability distribution for every variable. However, we have no information about correlations of the variables. Now, suppose that we know variables values for some sample of the objects. What is probability that this sample corresponds to some object? It is a typical problem of recognition over a condition of incomplete information.

Let us consider the simplest case when no correlations exist between variables. In this case, the Naive Bayes model is an exact solution of the problem. We prove in this paper that for the case that we know nothing about correlation the Naive Bayes model is not exact, but *optimal* solution in some sense. More detailed, we prove that the Naive Bayes model gives minimal mean error over all possible models of correlation. We suppose in this proof that all correlations models have the equal probability. We think that this result can explain the described above mysterious optimality of Naive Bayes.

The paper is organized as following. In section II we give exact mathematical definition of the problem for two variables and two objects. In section III we define our notations. In section IV we give generic form of conditional probability for all possible correlations of our variables. In section V we define the restrictions of the functions describing the correlations. In section VI we give the definition a distance between two probability(correlation) models. In section VII we find restrains for our basic functions. In the section VIII we solve our main problem we prove optimality of the Naive Bayes model for uniform distribution of all possible correlations. In the section IX we find mean error between the Naive Bayes model and an actual model for uniform distribution of all possible correlations. In section X we consider the case more than two variables and objects. The last section is conclusions.

II. STATEMENT OF THE PROBLEM.

Let A be a random variable, with values in set $0, 1$. Assume that the *apriori* probability $P(A) = P(A = 1)$

is known and denote it by θ. Let X_1, X_2 be two random variables, with values in some set, say $]-\infty;+\infty[$. We are given the following information:$X_1 = x_1$ and $X_2 = x_2$ (obtained though measurement). Furthermore, we have two systems - "classifiers" , which given x_1 and x_2 produce:

$$P(A = 1/X_1 = x_1) = P(A/x_1) \doteq \alpha \tag{1}$$

$$P(A = 1/X_2 = x_2) = P(A/x_2) \doteq \beta \tag{2}$$

We wish to estimate the probability $P(A = 1/X_1 = x_1, X_2 = x_2) = P(A/x_1, x_2)$ in terms of α, β and θ. More specifically we wish to find a function $\Gamma_{opt}(\alpha, \beta, \theta)$ which on the average is the best approximation for $P(A/x_1, x_2)$ in a sense to be defined explicitly in the sequel (see FIG. 1.).

III. NOTATION AND PRELIMINARIES

$\rho_{X_1,X_2}(x_1, x_2)$ - joint PDF(probability density function) of X_1 and X_2. $\rho_{X_1,X_2/A}(x_1, x_2) \doteq h(x_1, x_2)$ -joint PDF of X_1 and X_2 , given $A = 1$. In terms of $h(x_1, x_2)$ and θ we may write $P(A/x_1, x_2)$ as follows:

$$P(A/x_1, x_2) = \frac{\theta h(x_1, x_2)}{\theta h(x_1, x_2) + (1-\theta)\overline{h}(x_1, x_2)} \tag{3}$$

where
$\overline{h}(x_1, x_2) \doteq \rho_{X_1,X_2/\overline{A}}(x_1, x_2)$- joint PDF of X_1 and X_2, given $A = 0$.
We have:

$$\rho_{X_1}(x_1) = \int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1, x_2)dx_2 \tag{4}$$

$$\rho_{X_2}(x_2) = \int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1, x_2)dx_1 \tag{5}$$

$$h_1(x_1) \doteq \rho_{X_1/A}(x_1) = \int_{-\infty}^{+\infty} h(x_1, x_2)dx_2 \tag{6}$$

$$h_2(x_2) \doteq \rho_{X_2/A}(x_2) = \int_{-\infty}^{+\infty} h(x_1, x_2)dx_1 \tag{7}$$

$$\overline{h}_1(x_1) \doteq \rho_{X_1/\overline{A}}(x_1) = \int_{-\infty}^{+\infty} \overline{h}(x_1, x_2)dx_2 \tag{8}$$

$$\overline{h}_2(x_2) \doteq \rho_{X_2/\overline{A}}(x_2) = \int_{-\infty}^{+\infty} \overline{h}(x_1, x_2)dx_1 \tag{9}$$

IV. GENERIC FORM OF $P(A/x_1, x_2)$

Define the function $g(x_1, x_2)$ and $\overline{g}((x_1, x_2)$

$$g(x_1, x_2) \doteq \frac{h(x_1, x_2)}{h_1(x_1)h_2(x_2)} \tag{10}$$

$$\overline{g}(x_1, x_2) \doteq \frac{\overline{h}(x_1, x_2)}{\overline{h}_1(x_1)\overline{h}_2(x_2)} \tag{11}$$

Note that if X_1 and X_2 are conditionally independent, i.e.

$$\begin{aligned} & h(x_1, x_2) = \rho_{X_1X_2/A}(x_1, x_2) = \rho_{X_1/A}(x_1)\rho_{X_2/A}(x_2) \\ = \quad & h_1(x_1)h_2(x_2) \end{aligned} \tag{12}$$

then

$$g(x_1, x_2) = \overline{g}(x_1, x_2) = 1 \tag{13}$$

Define the following *monotonously nondecreasing* probability distribution functions

$$H_1(x_1) \doteq \int_{-\infty}^{x_1} h_1(z)dz \tag{14}$$

$$H_2(x_2) \doteq \int_{-\infty}^{x_2} h_2(z)dz \tag{15}$$

$$\overline{H}_1(x_1) \doteq \int_{-\infty}^{x_1} \overline{h}_1(z)dz \tag{16}$$

$$\overline{H}_2(x_2) \doteq \int_{-\infty}^{x_2} \overline{h}_2(z)dz \tag{17}$$

Note 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 one could 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. This restriction will be shown to be superfluous in the sequel.) , there exist the inverse functions $H_1^{-1}(x_1), H_2^{-1}(x_2), \overline{H}_1^{-1}(x_1)$ and $\overline{H}_2^{-1}(x_2)$. We may therefore define:

$$J(a, b) \doteq g(H_1^{-1}(a), H_2^{-1}(b)) \tag{18}$$

$$\overline{J}(a, b) \doteq \overline{g}(\overline{H}_1^{-1}(a), \overline{H}_2^{-1}(b)) \tag{19}$$

For the sake of brevity we shall henceforth denote

$$J \doteq J(H_1(x_1), H_2(x_2)) = g(H_1^{-1}(H_1(x_1)), H_2^{-1}(H_2(x_2))) = g(x_1, x_2) \quad (20)$$

$$\overline{J} \doteq \overline{J}(\overline{H}_1(x_1), \overline{H}_2(x_2)) = \overline{g}(\overline{H}_1^{-1}(\overline{H}_1(x_1)), \overline{H}_2^{-1}(\overline{H}_2(x_2))) = \overline{g}(x_1, x_2) \quad (21)$$

By the definition

$$h(x_1, x_2) = Jh_1(x_1)h_2(x_2) \quad (22)$$

$$\overline{h}(x_1, x_2) = \overline{Jh}_1(x_1)\overline{h}_2(x_2) \quad (23)$$

We now have:

$$h_1(x_1) \doteq \rho_{X_1/A}(x_1) = \frac{\rho_{X_1}(x_1)P(A/x_1)}{P(A)} = \frac{\alpha\rho_{X_1}(x_1)}{\theta} \ . \quad (24)$$

$$h_2(x_2) \doteq \rho_{X_2/A}(x_2) = \frac{\rho_{X_2}(x_2)P(A/x_2)}{P(A)} = \frac{\beta\rho_{X_2}(x_2)}{\theta} \ . \quad (25)$$

$$\overline{h}_1(x_1) \doteq \rho_{X_1/\overline{A}}(x_1) = \frac{\rho_{X_1}(x_1)P(\overline{A}/x_1)}{P(\overline{A})} = \frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta} \quad (26)$$

$$\overline{h}_2(x_2) \doteq \rho_{X_2/\overline{A}}(x_2) = \frac{\rho_{X_2}(x_2)P(\overline{A}/x_2)}{P(\overline{A})} = \frac{(1-\alpha)\rho_{X_2}(x_2)}{1-\theta} \quad (27)$$

Hence, from (22),(23)

$$h(x_1, x_2) = J\frac{\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{\theta^2} \quad (28)$$

$$\overline{h}(x_1, x_2) = \overline{J}\frac{(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{(1-\theta)^2} \quad (29)$$

Now from (3)

$$P(A/x_1, x_2) = \frac{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2)}{\frac{J}{\theta}\alpha\beta\rho_{X_1}(x_1)\rho_{X_2}(x_2) + \frac{\overline{J}}{(1-\theta)}(1-\alpha)(1-\beta)\rho_{X_1}(x_1)\rho_{X_2}(x_2)} = \frac{\alpha\beta}{\alpha\beta + \frac{\overline{J}}{J}\frac{\theta}{1-\theta}(1-\alpha)(1-\beta)} \quad (30)$$

Note that in case of conditional independence $J = \overline{J} = 1$ and (30) becomes the exact solution $\Gamma(\alpha, \beta, \theta) = P(A/x_1, x_2)$.

V. RESTRICTIONS ON THE FUNCTIONS $J(a, b)$ AND $\overline{J}(a, b)$

We have

$$h_1(x_1) = \int_{-\infty}^{+\infty} J(H_1(x_1), H_2(x_2))h_1(x_1)h_2(x_2)dx_2 \ . \quad (31)$$

Hence

$$1 = \int_{-\infty}^{+\infty} J(H_1(x_1), H_2(x_2))h_2(x_2)dx_2 = \int_0^1 J(H_1(x_1), H_2(x_2))dH_2(x_2) \quad (32)$$

Thus, we have the following condition

$$\int_0^1 J(a, b)db = 1 \quad (33)$$

and analogously

$$\int_0^1 J(a, b)da = 1 \quad (34)$$

Similarly, we obtain:

$$\int_0^1 \overline{J}(a, b)da = 1$$
$$\int_0^1 \overline{J}(a, b)db = 1 \quad (35)$$

Obviously

$$J(a, b), \overline{J}(a, b) \geq 0 \ . \quad (36)$$

$$\int_0^1 \int_0^1 J(a, b)dadb = \int_0^1 \int_0^1 \overline{J}(a, b)dadb = 1 \ . \quad (37)$$

The set of all the solutions of (33),(34),(35),(36),(37) together with (30) determines the set of all possible realizations of $P(A/x_1, x_2)$.

An example of a solution of (33),(34) *and* (36),(37). Let $\rho(x)$ be a function such that $\rho(x) \geq 0$ and $\int_0^1 \rho(x)dx = 1$
Then

$$J(a, b) = \begin{cases} \rho(a-b) & , a \geq b \\ \rho(a-b+1) & , a < b \end{cases} \quad (38)$$

satisfies (33),(34) and (36),(37).

VI. 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:

$$||\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)|| \doteq \int\int -\infty^{+\infty}\rho_{X_1X_2}(x_1,x_2) [\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)]^2 dx_1dx_2 \quad (39)$$

Now we have from (22),(23) and (24),(25), (26),(27)

$$\rho_{X_1X_2}(x_1,x_2)=\theta h(x_1,x_2)+(1-\theta)\overline{h}(x_1,x_2)= \theta J h_1(x_1)h_2(x_2)+(1-\theta)\overline{J}\overline{h}_1(x_1)\overline{h}_2(x_2)= [\frac{J\alpha\beta}{\theta}+\frac{\overline{J}(1-\alpha)(1-\beta)}{(1-\theta)}]\rho_{X_1}(x_1)\rho_{X_2}(x_2) \quad (40)$$

$$\begin{aligned}&||\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)||\\ &=\int\int_{-\infty}^{+\infty}\rho_{X_1}(x_1)\rho_{X_2}(x_2)\\ &[\frac{J\alpha\beta}{\theta}+\frac{\overline{J}(1-\alpha)(1-\beta)}{(1-\theta)}]\\ &(\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2))^2dx_1dx_2\\ &=\int_0^1\int_0^1[\frac{J\alpha\beta}{\theta}+\frac{\overline{J}(1-\alpha)(1-\beta)}{(1-\theta)}]\\ &(\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2))^2dF_1(x_1)dF_2(x_2)\end{aligned} \quad (41)$$

where

$$F_1(x_1)=\int_{-\infty}^{x_1}\rho_{X_1}(z)dz \quad (42)$$

$$F_2(x_2)=\int_{-\infty}^{x_2}\rho_{X_2}(z)dz \quad (43)$$

VII. RESTRAINTS FOR BASIC FUNCTIONS

We will consider in further all functions with arguments $1\geq F1,F2\geq 0$, but not x_1,x_2. We have six function of $F1,F2$, that define (41): $J,\overline{J},H_1,H_2,\alpha,\beta$. Let us to write the other function by help these function and find restraints for these functions.

(i)

$$\alpha=P(A/x_1)=\theta h_1(x_1)/\rho_{X_1}(x_1)=\theta\frac{\frac{dH_1}{dx_1}}{\frac{dF_1}{dx_1}}=\theta\frac{dH_1}{dF_1} \quad (44)$$

By the same way

$$\beta=\theta\frac{dH_2}{dF_2} \quad (45)$$

We know that functions H_1,F_1,H_2,F_2 are cumulative distribution functions of x_1,x_2, correspondently. These functions are *monotonously nondecreasing* functions and changes 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 :

$$\begin{aligned}H_1(1)&=H_2(1)=1\\ H_1(0)&=H_2(0)=0\end{aligned} \quad (46)$$

$$0\leq\alpha=\theta\frac{dH_1}{dF_1},\beta=\theta\frac{dH_2}{dF_2}\leq 1 \quad (47)$$

$$0\leq\theta\leq 1 \quad (48)$$

(ii)

$$\begin{aligned}&\overline{H}_1(x_1)=\int_{-\infty}^{x_1}\overline{h}_1(x_1)=\int_{-\infty}^{x_1}\frac{(1-\alpha)\rho_{X_1}(x_1)}{1-\theta}dx_1=\\ &\frac{1}{1-\theta}\int_{-\infty}^{x_1}-\frac{\theta}{1-\theta}\int_{-\infty}^{x_1}\frac{\alpha\rho_{X_1}(x_1)}{\theta}dx_1\\ &=\frac{F_1}{1-\theta}-\frac{\theta}{1-\theta}H_1(x_1)\end{aligned} \quad (49)$$

By the same way

$$\overline{H}_2(x_2)=\frac{F_2}{1-\theta}-\frac{\theta}{1-\theta}H_2(x_2) \quad (50)$$

(iii)

$$\begin{aligned}&J(H_1(F_1),H_2(F_2)):\\ &J(H_1(F_1),H_2(F_2))\geq 0\\ &\int_0^1 J(a,b)db=1\\ &\int_0^1 J(a,b)da=1\end{aligned} \quad (51)$$

$$\begin{aligned}&\overline{J}(\overline{H}_1(F_1),\overline{H}_2(F_2)):\\ &\overline{J}(\overline{H}_1(F_1),\overline{H}_2(F_2))\geq 0\\ &\int_0^1 \overline{J}(a,b)db=1\\ &\int_0^1 \overline{J}(a,b)da=1\end{aligned} \quad (52)$$

(iv)

$$P(A/x_1,x_2)=\frac{\frac{J\alpha\beta}{\theta}}{\frac{J\alpha\beta}{\theta}+\frac{\overline{J}(1-\alpha)(1-\beta)}{1-\theta}} \tag{53}$$

VIII. OPTIMIZATION

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

$$min_{\Gamma(\alpha,\beta,\theta)}E[||\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)||]\longrightarrow\Gamma(\alpha,\beta,\theta) \tag{54}$$

where the expected value (or expectation, or mathematical expectation, or mean, or the first moment) $E[...]$ 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:

$$C\doteq\frac{J\alpha\beta}{\theta}+\frac{\overline{J}(1-\alpha)(1-\beta)}{(1-\theta)} \tag{55}$$

$$D\doteq\frac{J\alpha\beta}{\theta} \tag{56}$$

Then from(53) and (41)

$$\begin{aligned}&||\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)||=\\&\int_0^1\int_0^1 C(\Gamma(\alpha,\beta,\theta)-D/C)^2dF_1dF_2=\\&\int_0^1\int_0^1 dF_1dF_2[\frac{D^2}{C}+\Gamma^2(\alpha,\beta,\theta)C-2\Gamma(\alpha,\beta,\theta)D]\end{aligned} \tag{57}$$

Thus

$$\begin{aligned}&min_{\Gamma(\alpha,\beta,\theta)}E[||\Gamma(\alpha,\beta,\theta)-P(A/x_1,x_2)||]=\\&min_{\Gamma(\alpha,\beta,\theta)}E[\int_0^1\int_0^1 dF_1dF_2\\&[\frac{D^2}{C}+\Gamma^2(\alpha,\beta,\theta)C-2\Gamma(\alpha,\beta,\theta)D]]=\\&min_{\Gamma(\alpha,\beta,\theta)}E[\int_0^1\int_0^1 dF_1dF_2[\frac{D^2}{C}]]+\\&min_{\Gamma(\alpha,\beta,\theta)}E[\int_0^1\int_0^1 dF_1dF_2\\&[\Gamma^2(\alpha,\beta,\theta)C-2\Gamma(\alpha,\beta,\theta)D]]=\\&Const+min_{\Gamma(\alpha,\beta,\theta)}E[\int_0^1\int_0^1 dF_1dF_2\\&[\Gamma^2(\alpha,\beta,\theta)C-2\Gamma(\alpha,\beta,\theta)D]]\end{aligned} \tag{58}$$

It remains to calculate the expected value in (58).

We have by obvious assumptions

$$\begin{aligned}&\rho_{J,\overline{J},\alpha,\beta,H_1,H_2/F_1,F_2}(J,\overline{J},\alpha,\beta,H_1,H_2/F_1,F_2)=\\&\rho_{J/H_1,H_2}(J/H_1,H_2)\rho_{\overline{J}/\overline{H}_1,\overline{H}_2}(\overline{J}/\overline{H}_1,\overline{H}_2)\\&\rho_{\alpha/F_1}(\alpha/F_1)\rho_{H_1/\alpha,F_1}(H_1/\alpha,F_1)\\&\rho_{\beta/F_2}(\beta/F_2)\rho_{H_2/\beta,F_2}(H_2/\beta,F_2)\end{aligned} \tag{59}$$

A. Lemma 1

$$E[J(a,b)]=\int_0^{+\infty}\rho_{J(a,b)/a,b}(J(a,b)/a,b)J(a,b)dJ=1 \tag{60}$$

$$E[\overline{J}(a,b)]=\int_0^{+\infty}\rho_{\overline{J}(a,b)/a,b}(\overline{J}(a,b)/a,b)\overline{J}(a,b)d\overline{J}=1 \tag{61}$$

Proof:

Let us consider function: $\rho_{J(a,b)/a,b}$. Function $J(a,b)$ is defined on the square $0\le a,b\le 1$. Let us make sampling of function J on this square by its dividing on small squares (i,j) and define value of the function J_{ij} on every square i,j. Restraints for function J (***) can be written

$$J_{ij}\ge 0 \tag{62}$$

$$\frac{1}{N}\sum_{i=1}^{N}J_{ij}=1 \tag{63}$$

$$\frac{1}{N}\sum_{j=1}^{N}J_{ij}=1 \tag{64}$$

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

All matrixes (J_{ij}) that satisfy these conditions are equal probability. Let us define probability density function

$$\rho(J_{11},...,J_{ij},...,J_{NN}) \tag{65}$$

This density function must be symmetric with respect to transpositions lines and columns in matrix (J_{ij}) , because the density function has equal probability for all matrixes (J_{ij}) that satisfy the above conditions. Indeed, these conditions are also symmetric with respect to transpositions lines and columns in matrix (J_{ij}). From symmetry conditions that define this function (ρ) with respect to transpositions lines and columns in matrix (J_{ij})

we can conclude that this function (ρ) also doesn't transform with respect to such transpositions.

Let us consider function $\rho_{u/ij}(u/ij)$ which is a discrete version of the function $\rho_{J(a,b)/a,b}(J(a,b)/a,b)$:

$$\rho_{u/ij}(u/ij) = \int ... \int_0^{+\infty} \rho(J_{11}, ..., J_{nk}, ..., J_{ij} = u, ..., J_{NN}) \prod_{(lm)\neq(ij)} dJ_{lm} \quad (66)$$

Let us transpose lines and columns (J_{ij}) by such way that element J_{ij} will be replaced by element J_{nk} , the function $\rho(J_{11}, ...)$ will not be transform after it. So from previous equation we obtain

$$\rho_{u/ij}(u/ij) = \int ... \int_0^{+\infty} \rho(J_{11}, ..., J_{nk} = u, ..., J_{ij}, ..., J_{NN}) \prod_{(lm)\neq(nk)} dJ_{lm} = \rho_{u/nk}(u/nk) \quad (67)$$

From this equation we can conclude that $\rho_{u/ij}(u/ij)$ doesn't depend on ij so $\rho_{J/ab}(J/ab)$ doesn't depend on ab and

$$\rho_{J/ab}(J/ab) = \rho_J(J) \quad (68)$$

and

$$E[J(a,b)] = \int_0^{+\infty} \rho_J(J) J dJ = Const \quad (69)$$

from

$$\int_0^1 \int_0^1 J(a,b) da db = 1 \quad (70)$$

we can conclude that

$$\int_0^1 \int_0^1 E[J(a,b)] da db = 1 \quad (71)$$

So we can obtain that $Const = 1$ in Eq.(69).

B. Lemma 2

Probability distribution functions α and β do not dependent on F_1 and F_2.

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha) \quad (72)$$

$$\rho_{\beta/F_2}(\beta/F_2) = \rho_\beta(\beta) \quad (73)$$

Proof:

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

$$0 \leq \alpha_k \leq 1 \quad (74)$$

$$\frac{1}{N} \sum_{k=1}^{N} \alpha_k = \int_0^1 \theta \frac{dH_1(F_1)}{dF_1} dF_1 = \theta \quad (75)$$

All columns (α_k) that satisfy by this conditions are equal probability. Let us to consider respective function $\rho(\alpha_1, ..., \alpha_k, ..., \alpha_l, ..., \alpha_N)$. From symmetry conditions that define this function with respect to transpositions $\alpha_k \to \alpha_l$ function $\rho(\alpha_1, ..., \alpha_k, ..., \alpha_l, ..., \alpha_N)$ also doesn't transform with respect to such transpositions. So we can write

$$\rho_k(u) = \int_0^1 \rho(\alpha_1, ..., \alpha_k = u, ..., \alpha_l, ..., \alpha_N) \prod_{n\neq k} d\alpha_n = \int_0^1 \rho(\alpha_1, ..., \alpha_k, ..., \alpha_l = u, ..., \alpha_N) \prod_{n\neq l} d\alpha_n = \rho_l(u) \quad (76)$$

From this equation, we can conclude that function $\rho_{\alpha/F_1}(\alpha/F_1)$ doesn't depend on F_1.

$$\rho_{\alpha/F_1}(\alpha/F_1) = \rho_\alpha(\alpha) \quad (77)$$

From (59) we obtain

$$E[\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D] = \int_0^1 \int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta) d\alpha d\beta \int_0^1 \int_0^1 \rho_{H_1/\alpha,F_1}(H_1/\alpha, F_1)\rho_{H_2/\beta,F_2}(H_2/\beta, F_2) dH_1 dH_2 \int_0^\infty \int_0^\infty \rho_J(J)\rho_{\overline{J}}(\overline{J})[\Gamma^2(\alpha,\beta,\theta)[\frac{J\alpha\beta}{\theta} + \frac{\overline{J}(1-\alpha)(1-\beta)}{1-\theta}] - 2\Gamma(\alpha,\beta,\theta)\frac{J\alpha\beta}{\theta}] dJ d\overline{J} = \int_0^1 \int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta) d\alpha d\beta [\Gamma^2(\alpha,\beta,\theta)[\frac{E[J]\alpha\beta}{\theta} + \frac{E[\overline{J}](1-\alpha)(1-\beta)}{1-\theta}] - 2\Gamma(\alpha,\beta,\theta)\frac{E[J]\alpha\beta}{\theta}] \quad (78)$$

Let us define

$$\overline{C} = \frac{\alpha\beta}{\theta} + \frac{(1-\alpha)(1-\beta)}{1-\theta} \tag{79}$$

$$\overline{D} = \frac{\alpha\beta}{\theta} \tag{80}$$

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

$$E[\Gamma^2(\alpha,\beta,\theta)C - 2\Gamma(\alpha,\beta,\theta)D] = \int_0^1\int_0^1 [\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}]\rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta \tag{81}$$

It remain to find

$$min_{\Gamma(\alpha,\beta,\theta)} \int_0^1\int_0^1 dF_1 dF_2 \int_0^1\int_0^1 d\alpha d\beta \rho_\alpha(\alpha)\rho_\beta(\beta) [\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}] \tag{82}$$

Since

$$\rho_\alpha(\alpha)\rho_\beta(\beta) \geq 0 \tag{83}$$

if the expression in square brackets is minimized at each point then the whole integral in (82) is minimized. Thus, we may proceed as follows

$$\frac{\partial}{\partial\Gamma}[\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}] = 2\Gamma(\alpha,\beta,\theta)\overline{C} - 2\overline{D} = 0 \tag{84}$$

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

$$\Gamma_{opt}(\alpha,\beta,\theta) = \frac{\overline{D}}{\overline{C}} = \frac{\frac{\alpha\beta}{\theta}}{\frac{\alpha\beta}{\theta} + \frac{(1-\alpha)(1-\beta)}{1-\theta}} \tag{85}$$

IX. MEAN DISTANCE BEETWEN 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 (57) is

$$DIS = E[||\Gamma(\alpha,\beta,\theta) - P(A/x_1,x_2)||] = \int_0^1\int_0^1 \rho_\alpha(\alpha)\rho_\beta(\beta)d\alpha d\beta [\Gamma^2(\alpha,\beta,\theta)\overline{C} - 2\Gamma(\alpha,\beta,\theta)\overline{D}] + Const \tag{86}$$

where $Const$ in this equation is defined by

$$Const = E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1 dx_2] \tag{87}$$

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

$$Const \leq E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)P(A/x_1,x_2)dx_1dx_2] = E[\theta] = \theta \tag{88}$$

The second condition is

$$0 \leq E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2) [P(A/x_1,x_2) - \theta]^2 dx_1dx_2] = E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2) [P(A/x_1,x_2)^2 + \theta^2 - 2P(A/x_1,x_2)\theta]dx_1dx_2] = E[\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho_{X_1,X_2}(x_1,x_2)[P(A/x_1,x_2)]^2 dx_1dx_2] - \theta^2 \tag{89}$$

So from these two equations we can conclude

$$\theta^2 \leq Const \leq \theta \tag{90}$$

By next step we would like find function $\rho_\alpha(\alpha)$ ($\rho_\beta(\beta)$) in equation for DIS.

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

(i)

$$\int_0^1 \alpha(F_1)dF_1 = \theta \tag{91}$$

(ii)

$$0 \leq \alpha(F_1) \leq 1 \tag{92}$$

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

(i)

$$\frac{1}{N}\sum_{i=1}^{N} \alpha_i = \theta \tag{93}$$

(ii)

$$0 \leq \alpha_i \leq 1, i = 1, 2, ..., N \tag{94}$$

Let us define function $U(\alpha_{set})$ by next way

$$U(\alpha_{set}) = \begin{cases} \sum_{i=1}^{N} \alpha_i & for 0 \leq \alpha_i \leq 1 \\ +\infty & otherwise \end{cases} , i = 1, 2, ..., N \tag{95}$$

$$U(\alpha_{set}) = \sum_{i=1}^{N} U_i(\alpha_i) \tag{96}$$

$$U_i(\alpha_i) = \begin{cases} \alpha_i & for 0 \leq \alpha_i \leq 1 \\ +\infty & otherwise \end{cases} \tag{97}$$

Then function that satisfies equal probability distribution with considering restrictions (i),(ii) is

$$\rho_{\alpha_{set}}(\alpha_{set}) = \frac{1}{C}\delta(U(\alpha_{set}) - N\theta) \tag{98}$$

where δ - delta-function of Dirac.
Constant C define by

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_{set}) d\alpha_1 ... d\alpha_N = 1 \tag{99}$$

It can be proved (see each course of "Statistical mechanics"; transform from microcanonical to canonical distribution) that for $N \mapsto \infty$ distribution (98) is equal to next distribution:

$$\rho_{\alpha_{set}}(\alpha_{set}) = \frac{1}{Z} e^{-KU(\alpha_{set})} \tag{100}$$

where Z and K can be found from equations

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_{set}) d\alpha_1 ... d\alpha_N = 1 \tag{101}$$

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} U(\alpha_{set}) \rho_{\alpha_{set}}(\alpha_{set}) d\alpha_1 ... d\alpha_N = N\theta \tag{102}$$

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

$$\rho_\alpha(\alpha) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \rho_{\alpha_{set}}(\alpha_1, ..., \alpha_j = \alpha, ..., \alpha_N) \prod_{i=1, i \neq j}^{N} d\alpha_i = \frac{1}{D} e^{-KU_j(\alpha_J = \alpha)} \tag{103}$$

where

$$D^N = Z \tag{104}$$

From Eqs.(101),(102) we can find

$$\frac{1}{Z} = (\frac{K}{1 - e^{-K}})^N \tag{105}$$

$$\theta = \Lambda(K) \tag{106}$$

where $\Lambda(K)$ is decreasing function

$$\Lambda(K) = \begin{cases} 1 & for K = -\infty \\ 0 & for K = +\infty \\ 1/2 & for K = 0 \\ \frac{1}{K} - \frac{1}{e^K - 1} & otherwise \end{cases} \tag{107}$$

If K is root of Eq (refpor6) we can write from Eqs.(103),(104),(105),(106) for function $\rho_\alpha(\alpha)$:

$$\rho_\alpha(\alpha) = \begin{cases} \begin{cases} For & K = 0 \\ 1 & for 0 \leq \alpha \leq 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & K = +\infty \\ 2\delta(\alpha) & 0 \leq \alpha \leq 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & K = -\infty \\ 2\delta(\alpha - 1) & 0 \leq \alpha \leq 1 \\ 0 & otherwise \end{cases} \\ \begin{cases} For & otherwise K \\ \frac{1}{D} e^{-K\alpha} & 0 \leq \alpha \leq 1 \\ 0 & otherwise \end{cases} \end{cases} \tag{108}$$

where $2\int_0^1 \delta(\alpha - 1) = 2\int_0^1 \delta(\alpha) = 1$ and

$$\frac{1}{D} = \frac{K}{1 - e^{-K}} \tag{109}$$

X. THE CASE OF MORE THAN TWO VARIABLES A AND X

Let A be a random variable, with values in set $0, 1, ..., L$. Assume that the *apriori* probability $P(A = i)$ is known and denote it by θ_i, here $i = 1, ..., L$. Let $X_1, ..., X_K$ be two random variables, with values in some set, say $]-\infty; +\infty[$. We are given the following information:$X_1 = x_1$,...,$X_K = x_K$ (obtained though measurement). Furthermore, we have two systems - "classifiers", which given x_1 ,...,x_K produce:

$$P(A = i / X_j = x_j) \doteq \alpha_{ij} \tag{110}$$

We wish to estimate the probability $P(A = i/X_1 = x_1, ..., X_K = x_K)$ in terms of α_{ij} and θ_i. More specifically we wish to find a function $\Gamma_{opt,M}(\alpha_{ij}, \theta_i)$ which on the average is the best approximation for $P(A = M/x_1, ..., x_K)$. By the same way, that in case of two variables we can find that the $\Gamma_{opt,M}(\alpha_{ij}, \theta_i)$ defined by equation

$$\Gamma_{opt,M}(\alpha_{ij}, \theta_i) = \frac{(\prod_{j=1}^{K} \alpha_{Mj})/\theta_M^{K-1}}{\sum_{i=1}^{L}(\prod_{j=1}^{K} \alpha_{ij})/\theta_i^{K-1}} \quad (111)$$

We have evidential restraints for α_{ij},θ_i

$$0 \leq \alpha_{ij} \leq 1$$
$$\sum_{i=1}^{L} \alpha_{ij} = 1 \quad (112)$$

$$0 \leq \theta_i \leq 1$$
$$\sum_{i=1}^{L} \theta_i = 1 \quad (113)$$

XI. CONCLUSIONS

We proved successfully that the Naive Bayes model gives minimal mean error over uniform distribution of all possible correlation between characteristic variables. This result can explain the described above mysterious optimality of Naive Bayes. We also found mean error that the Naive Bayes model gives for uniform distribution of all possible correlation.

Acknowledgments We would like to thank Alexsander Vardy and Romanov Alexey Nikolaevich for their help in creating this paper. We also would like to thank anonymous referee for very useful and clear remarks.

[1] Kupervasser O., Vardy A., Estimation of the Probability in the System of "Classifiers",(2002) arXiv:cs/0202020v1,http://arxiv.org/abs/cs/0202020v1

[2] Raymer M. L., Doom T. E.,Kuhn L. A.,Punch W. F., Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm, IEEE Transactions on Systems, Man, and Cybernetics, **33B**, 802 (2003)

[3] Domingos, P., and Pazzani, M., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning **29**, 103 (1997)

[4] Zhang H., The Optimality of Naive Bayes, In FLAIRS Conference (2004) http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf

[5] Kuncheva L.I., On the optimality of Naive Bayes with dependent binary features, Pattern Recognition Letters, **27**,830 (2006)

[6] Landau L.D., Lifshitz E.M., *Statistical Physics*, **5**,Elsevier Science Technology, United Kingdom, (1996)

[7] Pospelov D.A., *Iskustvennyj intelekt* (artificial intelligence), Handbook,"Radio i svjaz", Moskva, (1990)

[8] Ventsel, E. S. *Teorija verojatnostej* (Probability Theory in Russian), Nauka , Moscow (1969)

Figures Legends

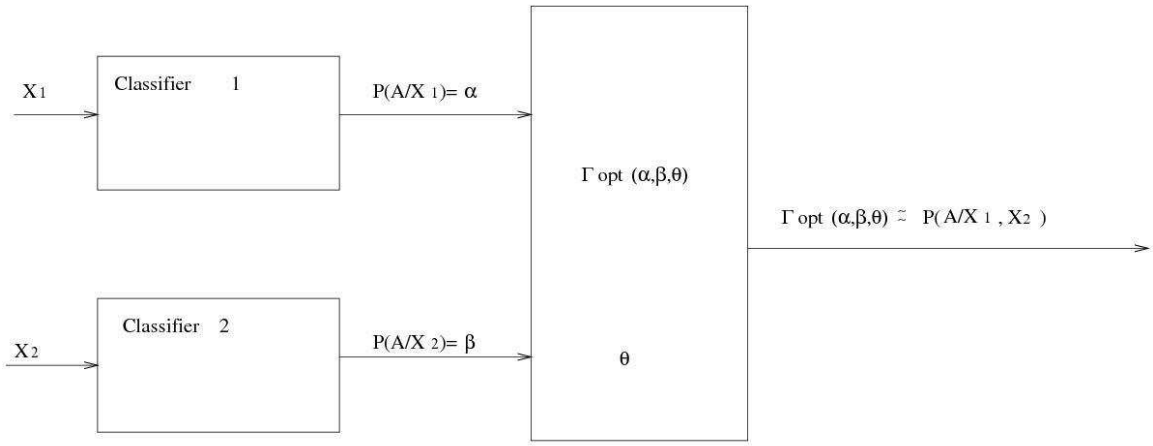

FIG. 1: Function $\Gamma(\alpha, \beta, \theta) : [0, 1]^3 \mapsto [0, 1]$