Mussa *et al. Journal of Cheminformatics* 2013, **5**:37
http://www.jcheminf.com/content/5/1/37

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Full "Laplacianised" posterior naive Bayesian algorithm

Hamse Y Mussa[1*], John BO Mitchell[2] and Robert C Glen[1]

Abstract

Background: In the last decade the standard Naive Bayes (SNB) algorithm has been widely employed in multi–class classification problems in cheminformatics. This popularity is mainly due to the fact that the algorithm is simple to implement and in many cases yields respectable classification results. Using clever heuristic arguments "anchored" by insightful cheminformatics knowledge, Xia *et al.* have simplified the SNB algorithm further and termed it the Laplacian Corrected Modified Naive Bayes (LCMNB) approach, which has been widely used in cheminformatics since its publication.
In this note we mathematically illustrate the conditions under which Xia *et al.*'s simplification holds. It is our hope that this clarification could help Naive Bayes practitioners in deciding when it is appropriate to employ the LCMNB algorithm to classify large chemical datasets.

Results: A general formulation that subsumes the simplified Naive Bayes version is presented. Unlike the widely used NB method, the Standard Naive Bayes description presented in this work is discriminative (not generative) in nature, which may lead to possible further applications of the SNB method.

Conclusions: Starting from a standard Naive Bayes (SNB) algorithm, we have derived mathematically the relationship between Xia *et al.*'s ingenious, but heuristic algorithm, and the SNB approach. We have also demonstrated the conditions under which Xia *et al.*'s crucial assumptions hold. We therefore hope that the new insight and recommendations provided can be found useful by the cheminformatics community.

Keywords: Naive Bayes, Laplacian Corrected Modified Naive Bayes, Classifications, Cheminformatics

Background

Broadly speaking there are two conceptually different ways to solve statistical problems: the frequentist and the Bayesian approaches. On the pros and cons of each method there are numerous excellent review articles and text books, such as the recent book by Murphy [1]. Unlike the frequentist approach, in the Bayesian approach any a priori knowledge about the probability distribution function that one assumes might have generated the given data (in the first place) can be taken into account when estimating this distribution function from the data at hand. If the data are noise–free and "complete", the role of the a priori information in estimating the distribution function diminishes drastically. However, the a priori information can be crucial when the data are noisy and sparse. The latter scenario is typical in realistic large chemical datasets, which, arguably, makes Bayesian based statistics a powerful data analysis tool.

Unfortunately, Bayesian statistics in its fullest form is not computationally feasible in realistic cheminformatics data analyses. However, in recent years, a simplified version of the Bayesian approach, which is commonly known as the "Naive" Bayesian algorithm, has been found to be a useful classification tool in multi–class classification problems in cheminformactics. To this end a Naive Bayesian classifier is built on binary descriptor space. The descriptors/features x_j, representing the compounds to be classified, assume binary values 0 or 1, where $(j = 1, 2, ..., L)$ and L can typically be more than 1,000. Thus for some cheminformatics practitioners even the Naive Bayesian algorithm in its standard form is computationally prohibitive when the dataset is large. In this regard, Xia *et al.* [2] proposed a simpler version of

*Correspondence: hym21@cam.ac.uk
[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK
Full list of author information is available at the end of the article

the standard Naive Bayesian algorithm, albeit for binary classification problems; slight variants of this algorithm for multi–class classification can also be found in [3,4]. According to Rogers *et al.* [5], Rogers being a co–author of the work presented in [2], "the standard Naive Bayes was modified by considering only the effect of the presence of a feature and not its absence". There are also a few more noticeable aspects of this proposed simplification: (a) the authors cleverly estimate directly – albeit heuristically – the a posteriori class probability for the present feature; (b) these authors (rather ingeniously) incorporate a Laplacian–correction into the estimated posterior class probability; and (c) the authors deem absent features not discriminating enough and therefore discard their contributions to the estimation of the posterior class. More than anything else it is this omission of the absent features from the Standard Naive Bayes (SNB) algorithm that makes Xia *et al.*'s proposed Naive Bayes Algorithm, termed Laplacian Corrected Modified Naive Bayes (LCMNB), (and its variants by different groups) computationally fast.

It is these three points, (a), (b) and (c), that we expound on in a mathematical setting to demonstrate under which conditions they hold – not only in an abstract sense, but also in the practical sense for a NB practitioner to make an informed decision as to when it is appropriate to employ SNB or LCMNB, in the cheminformatics context.

Methods

Naive Bayes

From Bayes' theorem recall that [6]:

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x})} \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_L)$ and ω_i denote the feature vectors and class labels, respectively; x_j and L being as described before, whereas i is just an index for the class labels. The terms $p(\omega_i|\mathbf{x})$, $p(\mathbf{x}|\omega_i)$, $p(\omega_i)$, and $p(\mathbf{x})$ refer to the posterior probability for ω_i given $\mathbf{x}$, the descriptor vector distribution conditioned on class ω_i, the a priori probability of class ω_i occurring, and the descriptor vector density function, respectively – for more details, see ref. [3,4,6].

The left hand side of Eq. 1 can be expressed as follows [1,7]

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(\omega_i|x_1, x_2, ..., x_L)}{p(\omega_i)} \tag{2}$$

By virtue of Bayes' theorem $p(\omega_i|x_1, x_2, ..., x_L)$ can be rewritten as

$$p(\omega_i|x_1, x_2, ..., x_L) = \frac{p(x_1, x_2, ..., x_L|\omega_i)p(\omega_i)}{p(x_1, x_2, ..., x_L)} \tag{3}$$

which in turn allows us to rewrite Eq. 2 as

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(\omega_i)p(x_1, x_2, ..., x_L|\omega_i)}{p(\omega_i)p(x_1, x_2, ..., x_L)} = \frac{p(x_1, x_2, ..., x_L|\omega_i)}{p(x_1, x_2, ..., x_L)} \tag{4}$$

Making use of the chain rule of probability [1,8], we can express $p(x_1, x_2, ..., x_L|\omega_i)$ as

$$p(x_1, x_2, ..., x_L|\omega_i) = p(x_1|\omega_i)p(x_2|\omega_i, x_1)..... \\ p(x_L|\omega_i, x_1, x_2, ..., x_{L-1}) \tag{5}$$

Plugging the right hand side of the equation above into Eq. 4 results in

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(x_1|\omega_i)p(x_2|\omega_i, x_1).....p(x_L|\omega_i, x_1, x_2, ..., x_{L-1})}{p(x_1, x_2, ..., x_L)} \tag{6}$$

In practice, it is extremely difficult to estimate $p(\omega_i|\mathbf{x})$ or $p(\mathbf{x}|\omega_i)$. This reality inevitably forces one to make concessions over the degree of accuracy the estimated $p(\omega_i|\mathbf{x})$ or $p(\mathbf{x}|\omega_i)$ can deliver. One widely employed scheme to obtain these probability distributions with compromised accuracy is to assume that individual descriptors x_j, $j = 1, 2, ..., L$, are independent conditional on ω_i. It is this naive assumption of independence among features to which the term "Naive" in "Naive Bayesian" refers.

Under this naive assumption, in Eq. 6, $p(x_2|\omega_i) = p(x_2|\omega_i, x_1)$, $p(x_3|\omega_i) = p(x_3|\omega_i, x_1, x_2)$,...,$p(x_L|\omega_i) = p(x_L|\omega_i, x_1, x_2..., x_{L-1})$. Thus, Eq. 6 modifies to

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(x_1|\omega_i)p(x_2|\omega_i)...p(x_L|\omega_i)}{p(x_1, x_2, ..., x_L)} \tag{7}$$

Multiplying top and bottom of Eq. 7 by $p^L(\omega_i)\Pi_{j=1}^L p(x_j)$ yields

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p^L(\omega_i)p(x_1|\omega_i)p(x_2|\omega_i)...p(x_L|\omega_i)\Pi_{j=1}^L p(x_j)}{p^L(\omega_i)p(x_1, x_2, ..., x_L)\Pi_{j=1}^L p(x_j)} \tag{8}$$

$$= \frac{p^L(\omega_i)p(x_1|\omega_i)p(x_2|\omega_i)...p(x_L|\omega_i)\Pi_{j=1}^L p(x_j)}{\Pi_{j=1}^L p(x_j)p^L(\omega_i)p(x_1, x_2, ..., x_L)} \tag{9}$$

Then making use of the fact that $p(\omega_i|x_1) = \frac{p(\omega_i)p(x_1|\omega_i)}{p(x_1)}$, $p(\omega_i|x_2) = \frac{p(\omega_i)p(x_2|\omega_i)}{p(x_2)}$,..., $p(\omega_i|x_L) = \frac{p(\omega_i)p(x_L|\omega_i)}{p(x_1)}$, we can rewrite Eq. 9 as

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{p(\omega_i|x_1)p(\omega_i|x_2)...p(\omega_i|x_L)\Pi_{j=1}^L p(x_j)}{p^L(\omega_i)p(x_1, x_2, ..., x_L)} \tag{10}$$

or more compactly as

$$\frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \frac{\Pi_{j=1}^L p(\omega_i|x_j)}{p^L(\omega_i)} \times \frac{\Pi_{j=1}^L p(x_j)}{p(x_1, x_2, ..., x_L)} \tag{11}$$

Clearly $\frac{\Pi_{j=1}^{L} p(x_j)}{p(x_1,x_2,\ldots,x_L)}$ is common to all classes and therefore plays no role in classification. Thus, in practice (in the Naive Bayes context with which this work is concerned) one is required to estimate $p(\omega_i|x_j)$ and $p(\omega_i)$.

Since generative approaches can be informative and "simpler" than their discriminative counterparts [9], we make use of Bayes' theorem again, i.e., $p(\omega_i|x_j) = \frac{p(\omega_i)p(x_j|\omega_i)}{p(x_j)}$ and then estimate $p(\omega_i|x_j)$ through $\frac{p(\omega_i)p(x_j|\omega_i)}{p(x_j)}$, where $p(x_j) = \sum_{i=1}^{C} p(x_j|\omega_i)p(\omega_i)$ with C referring to the number of classes. $p(\omega_i)$ denotes the a priori class probability, which is relatively easy to estimate. Thus, in our Bayesian context, the estimation of $p(\omega_j|x_j)$ boils down in practice to estimating $p(x_j|\omega_i)$.

Estimation of $p(x_j|\omega_i)$, with $x_j = 1$ *and* 0

$p(x_j|\omega_i)$ can be estimated using the given data and assuming a *Beta* distribution as an a priori distribution for $p(x_j|\omega_i)$ [10]. (There are other possible prior distributions from which one can choose, but we select the *Beta* distribution for reasons that will transpire later). As described in Appedix **A**, a *Beta* a priori distribution $Beta(\alpha_i, \beta_i)$ for $p(x_j|\omega_i)$ results in a $p(x_j|\omega_i)$ estimator in the form [11]:

$$p(x_j = 1|\omega_i) = \frac{N_{ij} + \alpha_i}{N_{\omega_i} + \beta_i + \alpha_i} \tag{12}$$

and of course

$$p(x_j = 0|\omega_i) = 1 - \frac{N_{ij} + \alpha_i}{N_{\omega_i} + \beta_i + \alpha_i} \tag{13}$$

where N_{ω_i} and N_{ij}, respectively, denote the number compounds in class ω_i, and number of compounds in this class with descriptor x_j assuming value 1. β_i and α_i are *Beta* distribution hyper–parameters per class and the valid range of values that these hyper–parameter can assume are as defined in Appendix **A**. When α_i and β_i equal 1, α_i and $\beta_i + \alpha_i$ in Eqs. 12–13 can be viewed as a "Laplacian correction".

Results and discussion

Estimation of $p(\omega_i|x_j = 1)$ and $p(\omega_i|x_j = 0)$

Estimation of $p(\omega_i|x_j = 1)$: In Our Approach

Remark 1. Assume that we have N chemical compounds (and their activity labels) available for training, where N_{ω_i} of these compounds belong to class ω_i.

Remark 2. Assume that the class a priori distribution is taken as $p(\omega_i) = \frac{N_{\omega_i}}{N}$, where $N_{\omega_i} >> \alpha_i + \beta_i$ (which is a valid assumption as found in any realistic large chemical dataset).

By virtue of Remark 1 and Eq. 12, the estimate of $p(\omega_i|x_j = 1)$ becomes

$$p(\omega_i|x_j = 1) = \frac{p(x_j = 1|\omega_i)p(\omega_i)}{p(x_j = 1)} = \frac{\frac{N_{ij}+\alpha_i}{N_{\omega_i}+\alpha_i+\beta_i} \times \frac{N_{\omega_i}}{N}}{\sum_{i=1}^{C} \frac{N_{ij}+\alpha_i}{N_{\omega_i}+\alpha_i+\beta_i} \times \frac{N_{\omega_i}}{N}} \tag{14}$$

(recall that $p(x_j) = \sum_{i=1}^{C} p(x_j|\omega_i)p(\omega_i)$).

Because of Remark 2, Eq. 14 can be simplified to

$$p(\omega_i|x_j = 1) = \frac{N_{ij} + \alpha_i}{\sum_{i=1}^{C} N_{ij} + \sum_{i=1}^{C} \alpha_i} = \frac{N_{ij} + \alpha_i}{N_j^+ + \sum_{i=1}^{C} \alpha_i} \tag{15}$$

where N_j^+ is the number of times x_j assumes the value 1.

Estimation of $p(\omega_i|x_j = 1)$: In Xia et al.'s Formulation

In the approach of Xia *et al.*, $p(\omega_i|x_j = 1)$ is estimated as

$$p(\omega_i|x_j = 1) = \frac{N_{ij} + A_i}{N_j^+ + K} \tag{16}$$

where K is as defined in Xia *et al.* and in their paper A_i is given as

$$A_i = p(\omega_i) \times K, \; with \; K = \sum_{i=1}^{C} A_i \; as \; \sum_{i=1}^{C} p(\omega_i) = 1$$

Eq. 16 constitutes what Xia *et al.* term "the Laplacian–Corrected Modified Naive Bayes (LCMNB)" estimator for $p(\omega_i|x_j = 1)$.

If α_i in Eq. 15 is set to A_i, Eq. 15 is exactly equivalent to Xia *et al.*'s estimator for $p(\omega_i|x_j = 1)$ as can be seen in Eq. 16.

We note in passing that in Xia *et al.*'s case, $C = 2$ and $p(\omega_2) = \frac{1}{K}$, which in their nomenclature denoted by $p(Active)$ – that is, $A_2 = 1$ while $A_1 = K - 1$.

Initially we employed the *Beta* a priori distribution for the class conditional distribution to ascertain the equivalence of Eqs. 15 and 16. Fortunately, however, we have ended up with the general equations (Eqs. 14 – 15) that not only encapsulate the LCMNB scheme of Xia *et al.*, but also subsume the other various variants of LCMNB, such as those discussed in Nidhi *et al.* and Nigsch *et al.*'s papers [3,4].

At any rate, let us proceed to the nub of this work: Identifying the conditions under which the LCMNB algorithm holds with respect to the SNB algorithm. But first we need to describe the estimation of $p(\omega_i|x_j = 0)$.

Estimation of $p(\omega_i|x_j = 0)$: In Our Approach

In regard to the case of $x_j = 0$, we make use of Remark 1, Remark 2 and Eq. 13, which yield an estimator for

$p(\omega_i|x_j = 0)$ as

$$p(\omega_i|x_j = 0) = \frac{N_{\omega_i} - (N_{ij} + \alpha_i)}{N - (N_j^+ + \sum_{i=1}^{C} \alpha_i)} \tag{17}$$

Naive Bayes: scoring function from

For notational convenience let us denote $\frac{N_{ij}+\alpha_i}{N_j^+ + \sum_{i=1}^{C} \alpha_i}$ in Eq. 15 and $\frac{N_{\omega_i} - (N_{ij}+\alpha_i)}{N - (N_j^+ + \sum_{i=1}^{C} \alpha_i)}$ in Eq. 17 by ξ_{ij} and ν_{ij}, respectively.

Thus, $p(\omega_i|x_j = 1)$ and $p(\omega_i|x_j = 0)$ may be written more succinctly as $p(\omega_i|x_j) = \xi_{ij}^{x_j} \nu_{ij}^{(1-x_j)}$, which allows us to express Eq. 11 more compactly as

$$\begin{aligned} \frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} &= \frac{\Pi_j p(\omega_i|x_j)}{p^L(\omega_i)} \times \frac{\Pi_j p(x_j)}{p(x_1, x_2, ..., x_L)} \\ &= \frac{\Pi_j \xi_{ij}^{x_j} \nu_{ij}^{(1-x_j)}}{p^L(\omega_i)} \times \frac{\Pi_j p(x_j)}{p(x_1, x_2, ..., x_L)} \end{aligned} \tag{18}$$

Now we come to the core of this work, under which conditions does the LCMNB algorithm hold with respect to the SNB algorithm? Before we answer this question, we deem it instructive and more insightful to map Eq. 18 monotonically to a discriminant function, a "scoring function" (so to speak).

To this end, taking the logarithm of Eq. 18 results in

$$S_{\omega_i}(\mathbf{x}) = \ln \frac{p(\omega_i|\mathbf{x})}{p(\omega_i)} = \ln \frac{\Pi_j \xi_{ij}^{x_j} \nu_{ij}^{(1-x_j)}}{p^L(\omega_i)} \times \frac{\Pi_j p(x_j)}{p(x_1, x_2, ..., x_L)} \tag{19}$$

$$= \sum_j x_j \ln \xi_{ij} + \sum_j (1 - x_j) \ln \nu_{ij} - L \times \ln p(\omega_i) \tag{20}$$

$$+ \sum_j \ln p(x_j) - \ln p(x_1, x_2, ..., x_L)$$

Self–evidently, the term $\left[\sum_j \ln p(x_j) - \ln p(x_1, x_2, ..., x_L) \right]$ is common to all classes and therefore does not play any role in classifying a given new compound. In other words, for practical classification purposes we are only interested in class dependent terms, i.e.,

$$D_{\omega_i}(\mathbf{x}) = \sum_j x_j \ln \xi_{ij} - \ln p(\omega_i) + \sum_j (1 - x_j) \ln \nu_{ij} - (L-1) \times \ln p(\omega_i) \tag{21}$$

where $S_{\omega_i}(\mathbf{x}) = D_{\omega_i}(\mathbf{x}) + \left[\sum_j \ln p(x_j) - \ln p(x_1, x_2, ..., x_L) \right]$

Conditions

In Xia *et al.*'s approach, the LCMNB algorithm, is none other than $\sum_j x_j \ln \xi_{ij} - \ln p(\omega_i)$ in Eq. 21. This means that in Xia *et al.*'s scheme the contributions from the terms depending on $x_j = 0$ for a given class, i.e.,

$$\sum_j (1 - x_j) \ln \nu_{ij} - (L - 1) \times \ln p(\omega_i),\ \forall i,\ i = 1, 2, .., C \tag{22}$$

are discarded. To the best of our knowledge, neither in Xia *et al.* nor in any other paper on the LCMNB approach has it been demonstrated that (i) the contribution of Eq. 22 is zero, i.e.,

$$\sum_j (1 - x_j) \ln \nu_{ij} - (L - 1) \times \ln p(\omega_i) = 0,\ \forall i,\ i = 1, 2, .., C \tag{23}$$

equally, in these papers, it has not been shown that (ii)

$$\begin{aligned} |\sum_j x_j \ln \xi_{ij} - \ln p(\omega_i)| >> |\sum_j (1 - x_j) \ln \nu_{ij} - (L - 1) \\ \times \ln p(\omega_i)|,\ \forall i,\ i = 1, 2, .., C \end{aligned} \tag{24}$$

nor has it been established that (iii)

$$\sum_j (1 - x_j) \ln \nu_{ij} - (L-1) \times \ln p(\omega_i) = constant,\ \forall i,\ i = 1, 2, .., C \tag{25}$$

Thus unless one (or more) of the above – (i), (ii) and (iii) – is (are) met, the assumption on which the Modified Naive Bayesian algorithm is based is questionable and therefore its practitioners should pay attention to this discrepancy; clearly it is not justifiable to discard from the onset the contribution of $\sum_j (1 - x_j) \ln \nu_{ij} - (L - 1) \times \ln p(\omega_i)$ simply because features x_j are absent, i.e. $x_j = 0$.

For completeness, we consider also the case of the highly popular class prior distribution, $p(\omega_i) = \frac{1}{C}$, i.e. $p(\omega_1) = p(\omega_2) = ... = p(\omega_C)$. We hasten to add that this option was not included in the LCMNB scheme. At any rate, by simply repeating the arguments in the preceding sections, it is straightforward to show that one ends up with Eq. 21. In this scenario, though, $L \times \ln p(\omega_i)$ is common to all classes and therefore does not play a role in classifying a new compound, i.e., $D_{\omega_i}(\mathbf{x})$ reduces to

$$D_{\omega_i}(\mathbf{x}) = \sum_j x_j \ln \xi_{ij} + \sum_j (1 - x_j) \ln \nu_{ij} \tag{26}$$

Conclusions

Starting from a standard Naive Bayes (SNB) algorithm, we have derived mathematically the relationship between Xia *et al.*'s ingenious, but heuristic algorithm, and the standard Naive Bayes approach. We also describe the conditions on which Xia *et al.*'s crucial assumption – contributions from absent feature can be discarded – holds. It is our hope that, with this new insight, cheminformaticians may now be able to efficiently use the Modified version of the standard Naive Bayes algorithm, as proposed by Xia *et al.*, and subsequently by Nidhi *et al.* and Nigsch *et al.*

Appendix

Appendix A: Estimator of $p(x_j|\omega_i)$

Here we give for completeness the proof that a priori *Beta* distribution leads to Eqs. 12 and 13 in the text.

For bookkeeping:

ω_i: class label indexed by i, $i = 1, 2, ..., C$.
C: Number of classes.
N_{ω_i}: Number of samples in class ω_i.
N_{ij}: Number of samples in class ω_i with feature $x_j = 1$, $j = 1, 2, ..., L$.
L: Number of features.

We state from the onset, in the following derivation we follow closely the descriptions given in ref. [10]. We also note, for clarity's sake, in the following analyses we abuse notation and use x_{jk} for both the random variable and its realization.

In this work, $\mathbf{x} \in \{0, 1\}^L$, i.e. $x_j \in \{0, 1\}$ and suppose that x_j are independent Bernoulli random variables (and this is in fact the assumption made in the Naive Bayesian approach). Thus, in the Naive Bayesian setting $p(\mathbf{x}|\omega_i)$ can be given as

$$p(\mathbf{x}|\omega_i) = \Pi_{j=1}^{L} Ber(x_j|\mu_{ij}) = \Pi_{j=1}^{L} \mu_{ij}^{x_j}(1-\mu_{ij})^{1-x_j} \tag{27}$$

where μ_{ij} is an estimate for the conditional probability that feature j occurs in class ω_i, and is what we are trying to estimate given a set of compounds assumed to belong to class ω_i. (In our context, μ_{ij} is an estimator for $p(x_j|\omega_i)$, where $p(x_j|\omega_i)$ is as defined in the text.)

To estimate μ_{ij} in a Bayesian framework, we first view μ_{ij} as a random variable, then choose an "appropriate" prior and likelihood for the random variable μ_{ij}.

Let us suppose that our a priori knowledge about the random variable μ_{ij} indicates that μ_{ij} is described by a *Beta* distribution, i.e.,

$$\pi(\mu_{ij}) = \frac{1}{B(\alpha_i, \beta_i)} \mu_{ij}^{\alpha_i - 1}(1-\mu_{ij})^{\beta_i - 1}, 0 \le \mu_{ij} \le 1, \alpha_i, \beta_i > 0,\ i = 1, 2, ...C \tag{28}$$

where $B(\alpha_i, \beta_i)$ ensures that the *Beta* distribution is normalised

Using the Bayes' theorem, then the posterior probability for μ_{ij} on the training data can be given by

$$\pi(\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}) = \frac{f(x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}|\mu_{ij})\pi(\mu_{ij})}{\int_0^1 f(x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}|\mu_{ij})\pi(\mu_{ij})d\mu_{ij}} \tag{29}$$

where $f(x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}|\mu_{ij})$ refers to the likelihood, and x_{j1}, x_{j2},... and $x_{jN_{\omega_i}}$ denote the j^{th} feature of the N_{ω_i} samples/compounds from class ω_i. As the samples are assumed independent, then $f(x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}|\mu_{ij})$ becomes $\Pi_{k=1}^{N_{\omega_i}} f(x_{jk}|\mu_{ij}) = \Pi_{k=1}^{N_{\omega_i}} \mu_{ij}^{x_{jk}}(1-\mu_{ij})^{1-x_{jk}}$, *i.e.*

$$\Pi_{k=1}^{N_{\omega_i}} f(x_{jk}|\mu_{ij}) = \mu_{ij}^{\sum_{k=1}^{m} x_{jk}}(1-\mu_{ij})^{N_{\omega_i} - \sum_{k=1}^{N_{\omega_i}} x_{jk}} \tag{30}$$

Thus, the posterior $\pi(\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jM})$ in Eq. 29 modifies to

$$\pi(\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}) = \frac{\mu_{ij}^{\sum_{k=1}^{N_{\omega_i}} x_{jk}}(1-\mu_{ij})^{N_{\omega_i} - \sum_{k=1}^{N_{\omega_i}} x_{jk}}\pi(\mu_{ij})}{\int_0^1 \mu_{ij}^{\sum_{k=1}^{N_{\omega_i}} x_{jk}}(1-\mu_{ij})^{N_{\omega_i} - \sum_{k=1}^{N_{\omega_i}} x_{jk}}\pi(\mu_{ij})d\mu_{ij}} \tag{31}$$

i.e.,

$$\pi(\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}) \propto \mu_{ij}^{\sum_{k=1}^{N_{\omega_i}} x_{jk}}(1-\mu_{ij})^{N_{\omega_i} - \sum_{k=1}^{N_{\omega_i}} x_{jk}}\pi(\mu_{ij}) \tag{32}$$

$$= \mu_{ij}^{\sum_{k=1}^{N_{\omega_i}} x_{jk}}(1-\mu_{ij})^{N_{\omega_i} - \sum_{k=1}^{N_{\omega_i}} x_{jk}} \times \mu_{ij}^{\alpha_i - 1}(1-\mu_{ij})^{\beta_i - 1} \tag{33}$$

$$= \mu_{ij}^{N_{ij} + \alpha_i - 1}(1-\mu_{ij})^{N_{\omega_i} - N_{ij} + \beta_i - 1} \tag{34}$$

Clearly, in Eq. 34, the posterior density for μ_{ij} given the samples $x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}$ has the same form as the prior for μ_{ij} [11], i.e.,

$$\pi(\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}) = \frac{1}{B(N_{ij} + \alpha_i, N_{\omega_i} - N_{ij} + \beta_i)} \mu_{ij}^{N_{ij} + \alpha_i - 1}(1-\mu_{ij})^{N_{\omega_i} - N_{ij} + \beta_i - 1} \tag{35}$$

which is none other than another *Beta* distribution. This means that the Bayes estimator of μ_{ij}, which is the estimate we are interested in, is the mean of the posterior distribution obtained [11]:

$$E[\mu_{ij}|x_{j1}, x_{j2}, ..., x_{jN_{\omega_i}}] = \frac{N_{ij} + \alpha_i}{N_{\omega_i} + \alpha_i + \beta_i} \tag{36}$$

In other words,

$$p(x_j|\omega_i) = \frac{N_{ij} + \alpha_i}{N_{\omega_i} + \alpha_i + \beta_i} \tag{37}$$

QED.

An accessible description of the derivation of Eq. 37 can be found in ref. [10].

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HYM conceived the idea that constitutes the nub of the presented work. This author also carried out the bulk of the mathematical derivations. RCG contributed to the Bayesian aspect of the work. JBOM conceptually contributed to the derivation given in Appendix A. The three authors participated in drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are in debt to Dr Dave Rogers for his many useful comments on the original LCMNB approach, in particular for helping us understand more about the two–class LCMNB version.
Mussa and Glen would like to thank the Unilever Centre for Molecular Sciences Informatics for its support, whereas Mitchell would like to thank the Scottish Universities Life Sciences Alliance (SULSA).

Author details

[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK. [2]EaStCHEM School of Chemistry and Biomedical Sciences Research Complex, University of St Andrews, North Haugh, St Andrews, Scotland, KY16 9ST, UK.

Received: 24 May 2013 Accepted: 12 August 2013
Published: 23 August 2013

References

1. Murphy KP: *Machine Learning: A Probabilistic Perspective*. 1st edition, Chapters 5 and 6; see Chapter 10 for the chain rule. Cambridge, MA: MIT Press; 2012.

2. Xia X, Maliski EG, Gallant P, Rogers D: **Classification of kinase inhibitors using a Bayesian model.** *J Med Chem* 2004, **47:**4463–4470.
3. Nidhi, Glick M, Davies JW, Jenkins JL: **Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases.** *J Chem Inf Model* 2006, **46:**1124–1133.
4. Nigsch F, Bender A, Jenkins JL, Mitchell JBO: **Ligand-target prediction using winnow and naive Bayesian algorithms and the implications of overall performance statistics.** *J Chem Inf Model* 2008, **48:**2313–2325.
5. Rogers D, Brown RD, Hahn M: **Using extended–connectivity fingerprints with Laplacian-modified Bayesian analysis in high–throughput screening follow–up.** *J Biomol Screen* 2005, **10:**682–686.
6. Townsend JA, Glen RC, Mussa HY: **Note on naive Bayes based on binary descriptors in Cheminformatics.** *J Chem Inf Model* 2012, **52:**2494–2500.
7. Duda RO, Hart PE: *Pattern Classification and Scene Analysis*. 1st edition, Chapter 2. New York, NY: John Wiley & Sons, Ltd; 1973.
8. Koch RK: *Introduction to Bayesian Statistics*. 2nd edition. Berlin: Springer; 2007.
9. Bishop CM: *Pattern Recognition and Machine Learning*. 1st edition, Chapter 1. New York: Springer; 2006.
10. Ross SM: *Introduction to Probability and Statistics for Engineers and Scientist*. 1st edition, Section 5. New York: John Wiley & Sons; 1987.
11. Davidson AC: *Statistical Models (Cambridge Series in Statistical and Probabilistic Mathematics)*. 1st edition. Cambridge: Cambridge University Press; 2008.

doi:10.1186/1758-2946-5-37
Cite this article as: Mussa *et al.*: Full "Laplacianised" posterior naive Bayesian algorithm. *Journal of Cheminformatics* 2013 **5**:37.