Wallenius Naive Bayes

Enric Junqué de Fortuny
Faculty of Applied Economics
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
enric.junquedefortuny@uantwerpen.be

David Martens
Faculty of Applied Economics
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
david.martens@uantwerpen.be

Foster Provost
Information, Operations & Management Sciences
Stern School of Business
New York University
fprovost@stern.nyu.edu

Abstract

Traditional event models underlying naive Bayes classifiers assume probability distributions that are not appropriate for binary data generated by human behaviour. In this work, we develop a new event model, based on a somewhat forgotten distribution created by Kenneth Ted Wallenius in 1963. We show that it achieves superior performance using less data on a collection of Facebook datasets, where the task is to predict personality traits, based on likes.

1 Description of the data/context

MAny data-analysis settings involve interactions of many entities of one kind with many entities of another. Often we find that there are resource constraints on the nature of these interactions. For instance, consider a dataset containing ingredient shopping lists of restaurants. Although there is a huge variety of ingredients, any one restaurant will only order so many of them (i.e. there are only so many distinct types of dishes a chef will make). The end result of such interactions are very high-dimensional datasets (in terms of variables), with relatively scarce information on any one data point (instance). Factoring in external properties such as seasonal and geological availability of certain ingredients, correlations between ingredients, cost, etc. results in datasets that are impossible to model perfectly. The usual approach is to simply ignore these difficulties and do a best-effort heuristic approach in which you try to model the data using faulty (or naive) assumptions. One well-known example are naive Bayes models, which have been shown to work extremely well in many settings despite their obvious defects. We would like to argue that in some cases (namely the one described before), it might be worth to account for (at least some) of the characteristics of the data. The main motivating example for the development of the work described in this manuscript are datasets that are a direct result from human behaviour.

As humans, we are faced with choices every day throughout the course of our lives. At the same time, we are often limited in the number of choices that we can actually make due to our bounded "behavioural capital", resulting from constraints like time, resources, etc. [10]. It is therefore not inconceivable to think that the select decisions we make, follow certain personal maxims which reason prescribes to us [12]. In a similar vein, Bourdieux argued in "La Distinction" [1] that we define ourselves through the choices that we make and as such they serve as a mirror for our own

predilections [1]. In many applications, one is interested in predicting properties over humans based on such predilections (recent examples include fraud detection [14], churn prediction [17] and direct marketing [15]).

To date it is still impossible to directly measure the intrinsic drivers behind these choices due to the intangibility of the human mind. Fortunately, recent socio-technological evolutions have allowed us to measure (and thus to study) the choices people make at a never before seen scale. As such it has now become possible to study how we distinguish ourselves through these various decisions.

In our set-up, we are concerned with subjects of two particular groups (hereafter called *classes*; e.g., smokers vs. non-smokers). Each of the subjects is allowed to make various choices, one by one, based on her intrinsic drivers. A subject can choose an option only once and is limited in the number of choices she can make. Can we then, given a history of such choices, predict what class a subject belongs to? The hypothesis is that due to the within-class similarity in intrinsic drivers, we could indeed correlate behaviour to particular groups.

2 Existing literature

2.1 Mathematical formulation of the problem

In our two-class set-up, we model the choices of an individual i as a vector $\mathbf{x}_i$. Each of its entries $x_{i,j}$ is Boolean valued and represents whether the user has chosen an item j or not. The full dataset then consists of set of labelled examples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, where the label y_i represents the class of the individual (either 0 or 1). The question is then whether we can predict the class labels for not previously seen individuals.

2.2 Multinomial Naive Bayes

The main idea behind the multinomial event model is that each input sample is a result from a series of independent trials from a bag of features (items). For each item j, present in an input vector $\mathbf{x}_i$, an independent trial is undertaken from an underlying multinomial distribution (for a total of $|\mathbf{x}|$ trials), each time picking one feature from all possible features ($X_1 \dots X_m$) with replacement. The probability of one such trial (picking a feature X_j from all possible features) is written as $P(X_j = x|C = c)$ and is independent of j, i.e., $P(X_{j_1} = x|C = c) = P(X_{j_2} = x|C = c)$. The aggregated probability of seeing the input vector given that it belongs to a certain class can then be modelled as a multinomial:

$$P(\mathbf{x}_i|C = c) = P(|\mathbf{x}_i|) \cdot |\mathbf{x}_i|! \cdot \prod_{j=1}^{m} \frac{P(X_j = x_{i,j}|C = c)^{x_{i,j}}}{x_{i,j}!} \tag{1}$$

The parameters for this model are the probabilities for each feature. Assuming a Laplacian prior on the parameters θ, we become the Bayes-optimal (maximal likelihood) estimates of the parameters for the formulation in Eq. 1 [16]:

$$\begin{aligned} \hat{\theta}_{X_j=x_j|C=c} &= P(X_j = x_j|C = c; \hat{\theta}_c) \\ &= \frac{1 + \sum_{i=1}^{n} x_{i,j} P(C = c|\mathbf{x}_i)}{m + \sum_{j=1}^{m} \sum_{i=1}^{n} x_{i,j} P(C = c|\mathbf{x}_i)} \end{aligned}$$

In the case of binary data this becomes:

$$\hat{\theta}_{X_j=x_j|C=c} = \frac{1 + |X_j = 1 \wedge C = c|}{m + \sum_{k=1}^{m} |X_k = 1 \wedge C = c|} \tag{2}$$

One of the main advantages of this formulation is that it only requires two passes over the active elements for all of the parameters to be estimated. Unfortunately, multinomial naive Bayes is not well aligned with our problem setting though. The cause of this is two-fold a) it was originally

[1] We must caution that Bourdieux limited himself to social class. We argue that, just like Hesse's Steppenwolf [9], it could very well be that there are many latent dimensions that define the choices we make and thus our character as a whole.

designed to work with frequency information (as opposed to binary information) and b) it does not take into account the influence of previous choices. As we will see in Section 3.3, this can result in heavily skewed posterior probability estimates. Previous research has shown that it is possible to incorporate additional dependencies, but the price of such bona-fide probability distributions is an exponential increase in time as the number of dependencies grows [7, 6].

2.3 Multi-Variate Bernouilli Naive Bayes

The multi-variate Bernoulli event model for naive Bayes assumes that the data ought to be generated according to a Bernoulli process. The resulting class conditional probability then becomes

$$P(\mathbf{x}_i|C=c) = \prod_{j=1}^{m} (\theta_j)^{x_{i,j}} (1-\theta_j)^{(1-x_{i,j})} \tag{3a}$$

Where θ_j is class conditional probability of feature j appearing in $x_{i,j}$. Optimizing the log likelihood, equating to zero and taking a Laplacian prior, yields the following parameter estimates for binary data [10]:

$$\begin{aligned} \hat{\theta}_{X_j=1|C=c} &= P(X_j = x_j|C=c;\hat{\theta}_c) \\ &= \frac{1+|X_j=1 \wedge C=c|}{2+|C=c|} \end{aligned}$$

Just like the multinomial event model, the multi-variate event model is very fast to compute, but does not take into account influence of context or correlations of the choices by design. Each choice is treated as a coin-flip with a fixed probability, estimated on the training data. As we will see in Section 3.3, this results in even more skewed posterior probabilities than the multinomial event model for human behavioural datasets.

3 Wallenius Naive Bayes

In this section we will first describe the general intuition behind the Wallenius distribution which lies at the core of our event model, followed by some refinements, technical remarks and a comparison to the previously mentioned event models.

3.1 A tale of fish

Imagine fishing for a meal from a pond of fish with m types of fish. There are m_j fish of type j for a total of $N = \sum_{j=1}^{m} m_j$ fish. We are fishing for a meal of exactly n fish, one by one. It turns out, that some fish are slower swimmers than others and are thus easier to catch. We will encode this as the easiness w_j for a type of fish j. The chances of a fish being caught are proportional to its easiness w_j and inversely proportional to the total easiness of the pond we are fishing at. Moreover, we are particularly hungry and eat the fish immediately after catching them. As such, once they are caught, their destiny is sealed and they are forever removed from the pond.

The probability of your particular meal $\mathbf{x}$ of the day with x_j fish of type j can be represented by a multi-variate Wallenius' non-central hypergeometric distribution with parameters: meal $\mathbf{x}$, quantity of fish $\mathbf{m}$ and easiness of fish $\mathbf{w}$. But let us first take look at a toy example to capture the intuition behind the distribution (Figure 1).

Consider catching fish $\{A, B, C, D\}$ from a pond with one of each type of fish. For this example we will assume some fish are slower than others; more precisely: $w_A = 4, w_B = 3, w_C = 2, w_D = 1$. One particular three course meal order (C after B after A) admits to a probability of:

$$\begin{aligned} P(A \to B \to C) &= P(A) \cdot P(B|A) \cdot P(C|B,A) \\ &= \frac{4}{10} \cdot \frac{3}{6} \cdot \frac{2}{3} \end{aligned} \tag{4}$$

Let us assume that we do not actually observe the order, but just the fish that had been caught. In order to calculate the probability of such an observation, we would then need to add up the probabilities of all the possible course orderings. We can represent the set of all possible course

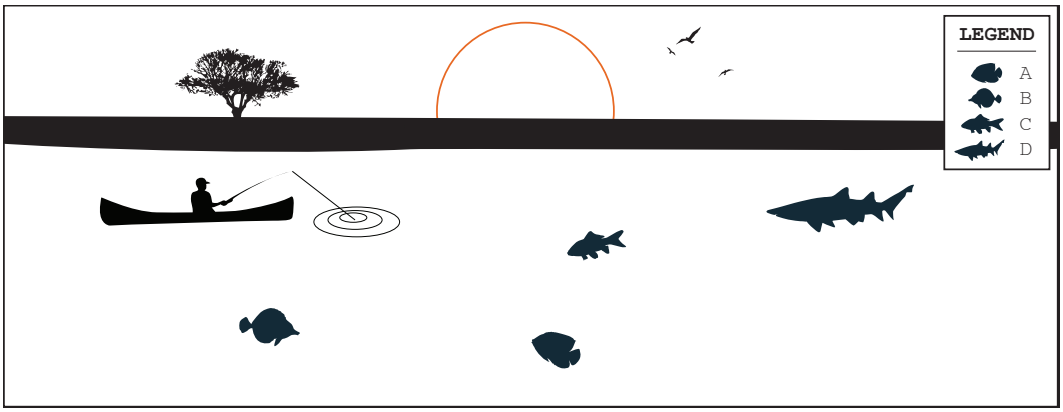

Figure 1: Jack is in the process of catching his meal of the day consisting of exactly three fish. The pond contains four fish with different catch probabilities, what is the probability of Jack catching combination $\{A, B, C\}$?

orderings by the permutation set $\sigma(\mathbf{x})$. Each $\mathbf{x}' \in \sigma(\mathbf{x})$ is one such possible ordered sequence of pickings (e.g., $A \to B \to C$ or $B \to A \to C$):

$$\begin{aligned}\mathbf{x}' &\sim x_{i_1} \to x_{i_2} \to \ldots \to x_{i_{|\mathbf{x}|}} \\ &\sim x'_1 \to x'_2 \to \ldots \to x'_{|\mathbf{x}|}\end{aligned}$$

Note that in this notation, $\mathbf{x}$ simply represents a sequence of events as opposed to the previously given vectorized binary format). Normalizing over the total number of possible permutations in the population then reveals the total probability of having caught this particular set of fish:

$$\begin{aligned}P(\mathbf{x}) &= P(\{A, B, C\}) \\ &= \frac{1}{Z}(P(A \to B \to C) + P(A \to C \to B) + P(B \to A \to C) + \\ &\quad P(B \to C \to A) + P(C \to A \to B) + P(C \to B \to A)) \\ &= \frac{1}{Z} \sum_{\mathbf{x}' \in \sigma(\mathbf{x})} P(\mathbf{x}') \end{aligned} \tag{5}$$

Generalizing, we get that:

$$P(\mathbf{x}, \mathbf{w}) = \frac{1}{Z} \sum_{\mathbf{x}' \in \sigma(\mathbf{x})} \prod_{j=1}^{|\mathbf{x}|} P(x'_j | x'_1 \ldots x'_{j-1}) \tag{6}$$

$$= \frac{1}{Z} \sum_{\mathbf{x}' \in \sigma(\mathbf{x})} \prod_{j=1}^{|\mathbf{x}|} \frac{w_{i_j}}{\sum_{k \geq j} w_{i_k}} \tag{7}$$

$$= \frac{1}{Z} \sum_{\mathbf{x}' \in \sigma(\mathbf{x})} \prod_{j=1}^{|\mathbf{x}|} \frac{w_j}{\sum_{k=1}^{m} w_k - \sum_{k<j} w_k} \tag{8}$$

$$Z = \binom{m}{|\mathbf{x}|}$$

In Eq. 6, we combine two elements from before: a) a multiplication over all elements in a sequence of draws to get the probability of that particular sequence (Eq.4) and b) a sum over all possible sequence permutations $\sigma(\mathbf{x})$ (Eq. 5). The normalizing constant Z is equal to the total number of possibilities in which one could have drawn a meal of size $|\mathbf{x}|$ from a pool containing m fish [2]. In Eq. 7 we then write the fractions explicitly by dividing the weight w_{i_j} of fish x'_j being caught at time

[2]Note that this particular normalizer assumes that all meals are of the same size, we can account for different assumptions by including appropriate normalizers.

j by the weights of the fish that are still left in the pool at that time in the sequence (i.e., any fish that we did not encounter yet and thus for which the index $k \geq j$. This is then expanded to arrive at the final formulation in Eq. 8.

This form is directly applicable, but the computations can take a long time. Using an efficient depth-first tree method with memory still requires about $\mathcal{O}(|\mathbf{x}|!)$ time, because we need to compute each of the permutations individually.

3.2 Wallenius' non-central Hypergeometric

3.2.1 General form

A more elegant solution was found by Wallenius in 1963 (and later generalized by Chesson in 1976 [2]) by looking at the problem as a Markov process and computing the stable points using backward Kolmogorov equations. This leads to the Wallenius distribution with probability mass function:

$$
\begin{aligned}
wall(\mathbf{x}, \mathbf{m}, \mathbf{w}) &= \Lambda(\mathbf{x}, \mathbf{m})\, I(\mathbf{x}, \mathbf{m}, \mathbf{w}) \\
\Lambda(\mathbf{x}, \mathbf{m}) &= \prod_{j=1}^{m} \binom{m_j}{x_j} \\
I(\mathbf{x}, \mathbf{m}, \mathbf{w}) &= \int_0^1 \prod_{j=1}^{m} \left(1 - t^{\frac{w_j}{s}}\right)^{x_j} dt \\
s &= \sum_{i=1}^{m} w_j
\end{aligned}
$$

Like before, $\mathbf{m}$ is a vector representation of the initial number of fish in the lake, and $\mathbf{w}$ of the weights of each of the fish. Here, we reintroduced the binary vector notation for $\mathbf{x}$, which is a vector of length c, containing a one ($x_j = 1$) if a fish of type j was caught and a zero ($x_j = 0$) if it survived. This expansion does not consider ordering information; this is one of its main strengths since it avoids the explicit summation over all of the permutations.

3.2.2 Binary form

Derivation Computing this integral is still quite hard and we would do well to simplify it as much as possible. In the case of binary variables ($\forall j : x_j \in \{0, 1\}$), we can rewrite the Wallenius distribution as:

$$
\begin{aligned}
wall(\mathbf{x}, \mathbf{m}, \mathbf{w}) &= \Lambda(\mathbf{x}, \mathbf{w})\, I(\mathbf{x}, \mathbf{m}, \mathbf{w}) \\
\Lambda(\mathbf{x}, \mathbf{m}) &= 1 \\
I(\mathbf{x}, \mathbf{m}, \mathbf{w}) &= \int_0^1 \prod_{j|x_j=1} \left(1 - t^{\frac{w_j}{s}}\right) dt \\
s &= \sum_{j=1}^{m} w_j(m_j - x_j) \\
&= \sum_{j|x_j=0} w_j
\end{aligned}
$$

The fractional exponent in the simplified form makes the integral very difficult to compute efficiently using typical numerical integration methods. Fortunately, we can transform the integrand to

a polynomial using the following variable substitution:

$$\begin{aligned} t &= u^s \\ dt &= s \cdot u^{s-1}\, du \\ wall(\mathbf{x}, \mathbf{m}, \mathbf{w}) &= \int_0^1 \prod_{j|x_j=1} \left(1 - t^{\frac{w_j}{s}}\right) dt \\ &= \int_0^1 \prod_{j|x_j=1} \left(1 - (u^s)^{\frac{w_j}{s}}\right) \cdot s \cdot u^{s-1}\, du \\ &= s \cdot \int_0^1 u^{s-1} \cdot \prod_{j|x_j=1} \left(1 - u^{w_j}\right) du \end{aligned} \tag{9}$$

Note that since $u = \sqrt[s]{t}$, the limits do not change. This accounts for all non-degenerate cases. The degenerate case of $s = 0$ (capture all fish) should of course return 1. Similar to other event models, there is a degeneration of the accuracy when features are unobserved since if $w_j = 0$ for any $j|x_j = 1$, this will cause $wall(\mathbf{x}, \mathbf{m}, \mathbf{w}) = 0$, which is not the desired behaviour. We can remedy this by adding a count to the weight vectors (i.e., each fish starts out with weight one). This corresponds to the prior belief that every feature is proportionally probable to its weight and results in a smoothing of the probabilities.

Since all w_j are integers and $s - 1$ is an integer, this is indeed a valid polynomial of order:

$$order = \sum_{j|x_j=1} w_j + (s - 1)$$

As a last simplification, we remark that the order of this polynomial can further be reduced by dividing the weights by their greatest common denominator.

Numerical evaluation A polynomial of integer degree $order$ can be computed exactly by the Gauss-Legendre polynomial quadrature method in $(order - 1)/2$ steps, which in our case becomes:

$$steps = \frac{1}{2} \sum_{j=1}^{m} w_j - 1$$

A approximate measure is given in [11], which states that for an integrand which has $2n = order$ continuous derivatives, the error can be bounded by:

$$erorr \leq \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^3} I^{(2n)}(\xi), \qquad a < \xi < b.$$

with I the integrand from before. In the end, this polynomial is still not trivial to calculate because of the huge possible order. It should be noted that other ways of calculating exist, but we will not further elaborate on them since this is out of the scope for of this manuscript (and a comprehensive overview is given in [8]). In our experiments we could deal with up to 1,000 prediction using 10,000 features in a reasonable amount of time on a low-end laptop.

3.2.3 Making predictions

As explained in the previous section, the final form given in Eq. 9 can be computed relatively easily. To actually predict class membership, one would have to calculate the class conditional probabilities for each class separately and compare them, resulting in a score that ranks input samples according to likelihood of class membership (revealed by applying the Bayes rule):

$$\begin{aligned} P(C = C_i|\mathbf{x}) &= \frac{P(C = C_i) \cdot P(\mathbf{x}|C = C_i)}{P(\mathbf{x})} \\ &\propto P(C = C_i) \cdot P(\mathbf{x}|C = C_i) \end{aligned} \tag{10}$$

$$\begin{aligned} Score(\mathbf{x}) &= \frac{P(C = C_0|\mathbf{x}) - P(C = C_1|\mathbf{x})}{P(C = C_0|\mathbf{x}) + P(C = C_1|\mathbf{x})} \\ &\propto \frac{P(C = C_0) \cdot P(\mathbf{x}|C = C_0) - P(C = C_1) \cdot P(\mathbf{x}|C = C_1)}{P(C = C_0) \cdot P(\mathbf{x}|C = C_0) + P(C = C_1) \cdot P(\mathbf{x}|C = C_1)} \end{aligned} \tag{11}$$

Here, we made the common assumption that each input sample is as likely as any other (revealing the likelihood in Eq. 10). The MLE for the prior class probability $P(C = C_i)$ can be calculated by looking at the fraction of samples belonging to class C_i in the total training set. The class conditional probabilities $P(\mathbf{x}|C = C_i)$ can be calculated using $wall(\mathbf{x}, \mathbf{m}, \mathbf{w}_i)$, where $\mathbf{w}_i$ is a weight vector for class i. These are estimated from the training set and require a linear pass through the active elements of the data. Making predictions requires two full computations of the Wallenius class-condition probability estimate (one for each class).

3.3 Comparison with other event models

We mentioned before that using the multinomial event model or the multi-variate Bernouilli event model can result into heavily skewed posterior probabilities. In this section we design an educational example to show where both of these might fail to yield well-calibrated probabilities.

Let us consider a scenario in which we try to predict gender based on movie-viewing history in a movie theatre where they only screen three movies: a blockbuster (M_1) and two niche movies (M_2 and M_3). We are given the movie viewing history for 100 male and 100 female visitors (as shown in Table 1).

Meet Sophie. Sophie has already seen the blockbuster (M_1) and is now faced with the choice between the predominantly male movie (M_2) and the mixed-gender movie (M_3), and thus decides to see M_3. Can we predict Sophie's gender given this choice?

Bernouilli's answer The conditional probability of Sophie seeing any of the movies is independent of her previous choices and comes down to a coin flip for each of the choices:

$$P(M_1, M_3|M) = \frac{90}{100} \cdot \frac{10}{100} \cdot \left(1 - \frac{10}{100}\right) = 8.10\%$$
$$P(M_1, M_3|F) = \frac{90}{100} \cdot \frac{10}{100} \cdot \left(1 - \frac{1}{100}\right) = 8.91\%$$

Due to the independence assumption, both genders are almost equally likely.

Multinomial answer The conditional probability of the events is independent of the context, but the probabilities should be normalized on a per class count:

$$P(M_1, M_3|M) = \frac{90}{110} \cdot \frac{10}{110} = 7.44\%$$
$$P(M_1, M_3|F) = \frac{90}{101} \cdot \frac{10}{101} = 8.82\%$$

The multinomial takes into account the fact that, faced with the choice between all three movies, a female would likely choose M_3 over M_2, but only in a very subtle way because the multinomial assumes that Sophie is given the choice of seeing all three of the movies every time. While this is true, the reality is that on average, the likelihood of Sophie repeatedly choosing M_1 is very low. This bias is reflected in the relatively marginal increase in probability for an otherwise very telling event.

Wallenius' answer Based on the sequence of events, the probabilities must be adapted to the context of each choice. We do not know the exact sequence, but we can average over all possible sequences:

$$P(M_1, M_3|M) = \binom{3}{2}^{-1} \cdot \left(\frac{90}{110} \cdot \frac{10}{110 - 90} + \frac{10}{110} \cdot \frac{90}{110 - 10}\right) = 16.36\%$$
$$P(M_1, M_3|F) = \binom{3}{2}^{-1} \cdot \left(\frac{90}{101} \cdot \frac{10}{101 - 90} + \frac{10}{101} \cdot \frac{90}{101 - 10}\right) = 30.27\%$$

Although there is still some bias due to the fact that we do not know the true sequence order, the answer is given with more certainty in this case (and this would be reflected in Eq. 11).

	Movie 1	Movie 2	Movie 3
F	90	1	10
M	90	10	10

Table 1: Artificial dataset of movie-watching behaviour of 100 male (M) and 100 female (F) subjects.

	Dataset	n	d	$order$	Predicted variable
[3]	Facebook Satisfaction	6,658	142,108	$\geq$ 1,437,932	satisfaction with life
[13]	Facebook IQ	6,377	134,420	$\geq$ 919,684	subject has high IQ
[13]	Facebook Gay	1,781	35,503	$\geq$ 232,130	exclusive same-sex interest for men
[4]	Facebook Smoking	3,746	95,186	$\geq$ 821,457	daily smoking behaviour
[4]	Facebook Alcohol	3,720	95,173	$\geq$ 816,909	daily alcohol drinking
[4]	Facebook Drugs	2,735	91,378	$\geq$ 609,070	daily drug using

Table 2: Overview of real life behavioural datasets.

4 Empirical evaluation

The example discussed in the previous section was constructed artificially and it stands to reason that real life datasets might behave differently. In this section, we compare the performance of all of the previously mentioned methods, based on human behavioural datasets.

4.1 Behavioural datasets

In order to test the method, we analysed the performance of Wallenius naive Bayes over real life datasets from Facebook (shown in Table 2) [13]. For each of these datasets, we have an input set of 'likes'. Within Facebook, a user can indicate whether or not she likes an item, we then use these likes as the input matrix to predict different personality traits. To assess the robustness and quality of prediction, we incrementally increase the number of features (likeable items) of the dataset. The predictive power is evaluated in terms of Area Under the ROC-curve [5] which, similar to the Gini coefficient, is an indicator for how well we rank randomly chosen samples form the dataset.

The resulting learning curves (Figure 2) show superior performance of Wallenius naive Bayes over five out of six datasets. To ensure the robustness of the results, each point in the curve is the result of a ten-fold randomized cross-testing procedure on hold-out test sets. Due to the previously mentioned computational complexity of Wallenius naive Bayes we were not able to run it on the full datasets yet (this is part of future research). Note however that Wallenius naive Bayes is sometimes able to outperform other event models when trained on the full dataset, even with limited information available.

The delta-plots (Figure 3) show the gap between Wallenius and the multi-variate and Bernouilli event model respectively. The plots show that further improvements are expected in most cases by further increasing the number of features, but most of the gain is attained in the initial data. Afterwards, the returns from using one model over teh other diminish due to the slowly decaying improvement rate.

4.2 Non-behavioural datasets

Given the promising results from the previous section, we also studied how Wallenius naive Bayes responds to non-behavioural datasets. It is well established that multinomial methods work very well for text-mining datasets and they clearly do not follow the premises of our event model (since Wallenius assumes that a document may only contain a term once). As such we included a subset of the famous RCV-1 dataset (called Dexter), where the predicted variable is whether the text is related to corporate acquisition or not. For the sake of completeness, we include both a binary version of the dataset, as well as one containing the actual term-frequencies.

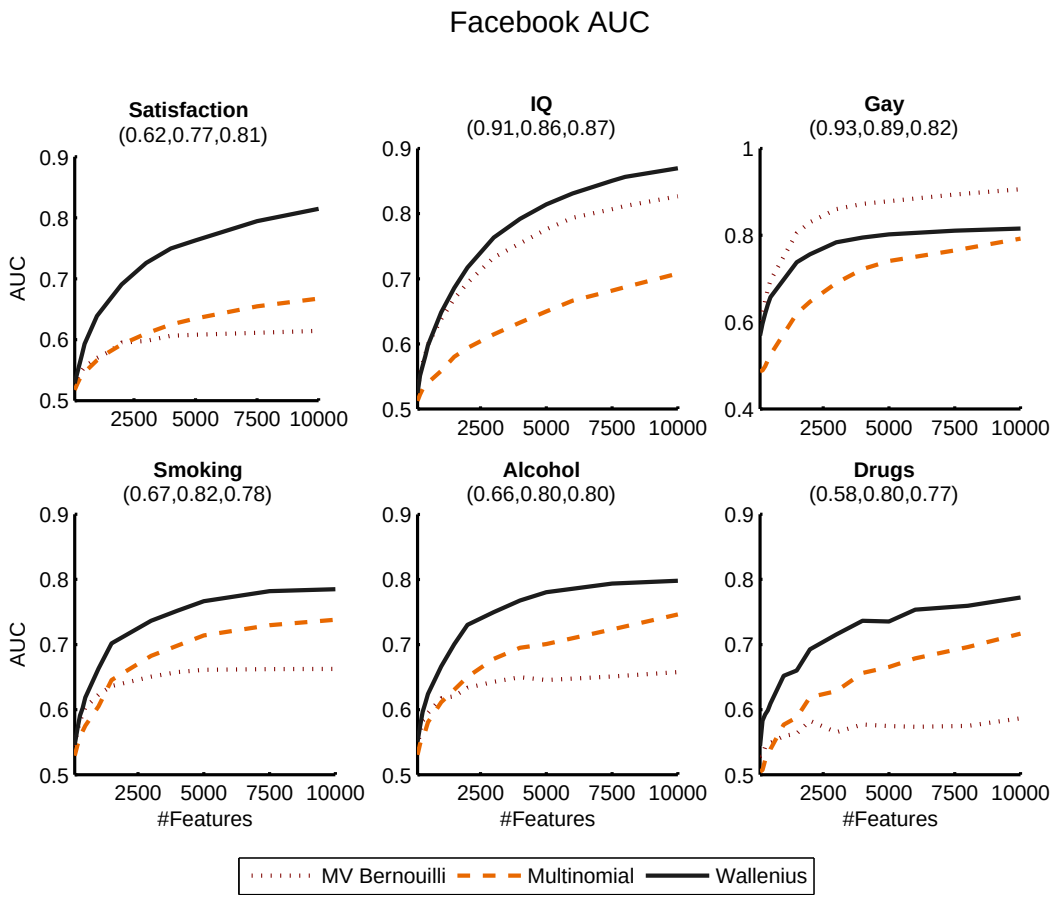

Figure 2: Experimental results for MV Bernoulli, the multinomial and the Wallenius event models for increasing number of features. Wallenius almost always performs Pareto-superior with respect to the other methods in terms of AUC. Also shown (under the title between brackets) are the AUC values using the full dataset for MV Bernouilli and Multinomial and the maximal AUC value of Wallenius in the learning curve.

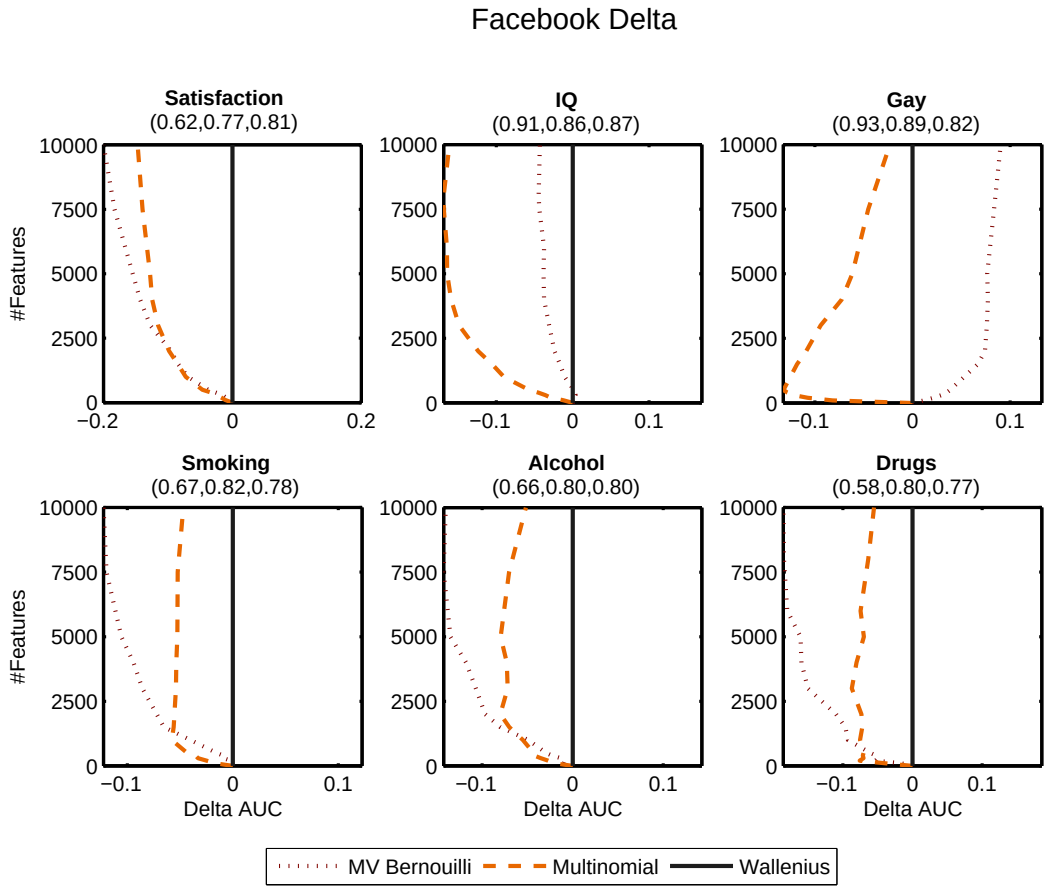

Figure 3: Delta-plot of the divergence between Wallenius and the multi-variate and Bernouilli event model respectively. The plots show that further improvements are expected in most cases by further increasing the number of features, although these do suffer from diminishing returns due to the slowly decaying improvement rate.

As can be seen in Figure 4, the Wallenius naive Bayes methods is quickly overtaken by both the binary and the regular multinomial naive Bayes. Interestingly, Wallenius naive Bayes does converge to its final performance with inclusion of a very limited number of features in a similar way as with the behavioural datasets.

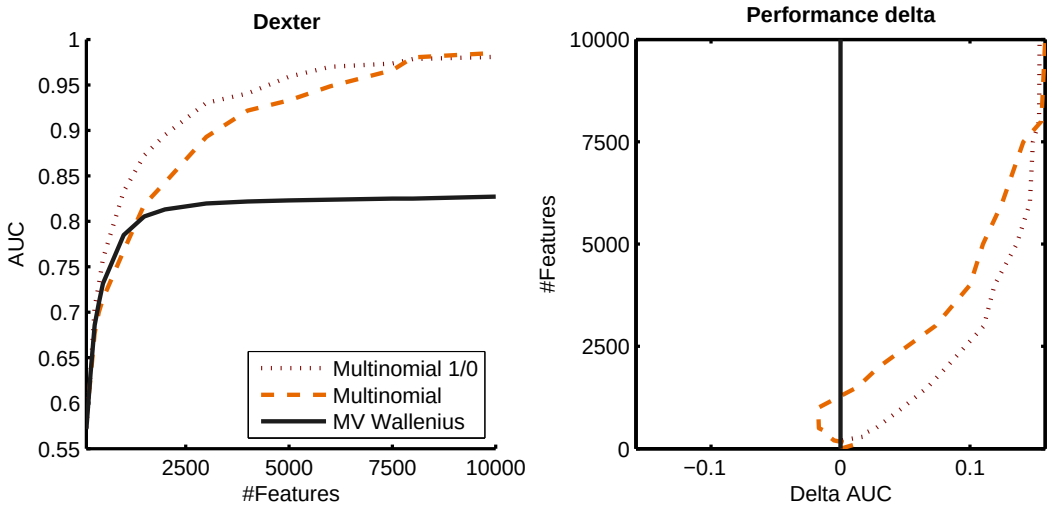

Figure 4: Comparison of multinomial naive Bayes and Wallenius naive Bayes for a text corpus prediction task. The multinomial event model clearly fares better in this case.

5 Conclusion

We developed a new variant of naive Bayes, based on the Wallenius distribution and have shown that it achieves better results with less data when working with human behavioural data. We emphasise the domain-specific scope of our results, as shown by the text-mining example in the experimental section. Indeed, for other domains, alternative event models might certainly still be a better choice. The work presented in this manuscript does show potential for the development of new event models to fit different scenarios (as opposed to blindly applying a traditional event model), one example being Fishers' non-central hypergeometrical distribution in non-competitive contexts. We hope that by showing the advantages of context-specific modelling, we can encourage other researchers to look into novel event models for different domains.

Given the promising results of Wallenius naive Bayes, we believe further research as to speeding up Wallenius naive Bayes might be worthwhile. Additionally, a natural extension of the currently presented model is to include a length prior to the distribution as well.

References

[1] P. Bourdieu. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press, 1984.

[2] J. Chesson. A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability*, 1976.

[3] E. Diener, R.A. Emmons, R.J. Larsen, and S. Griffin. The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 1985.

[4] J.-F. Etter, J. Le Houezec, and T.V. Perneger. A self-administered questionnaire to measure dependence on cigarettes: the cigarette dependence scale. *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, 28(2):359–70, February 2003.

[5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[6] P. Flach and N. Lachiche. Decomposing probability distributions on structured individuals. In *Work-in-Progress Reports of the 10th International Conference on Inductive Logic Programming*, pages 96–106, 2000.

[7] P. Flach and N. Lachiche. Naive Bayesian Classification of Structured Data. *Machine Learning*, 57(3):233–269, December 2004.

[8] A. Fog. Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics - Simulation and Computation*, 37(2):258–273, February 2008.

[9] H. Hesse. *Steppenwolf: A Novel*. Picador, 1927.

[10] E. Junqué de Fortuny, D. Martens, and F. Provost. Predictive Modeling with Big Data: Is Bigger Really Better? *Big Data*, (ahead of print), October 2013.

[11] D. Kahaner, C.B. Moler, and S. Nash. *Numerical methods and software*. Prentice Hall, 1989.

[12] I. Kant. *The Critique of Judgement (Part One, The Critique of Aesthetic Judgement)*. BiblioLife, 1790.

[13] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5802–5, April 2013.

[14] D. Martens, E. Junqué de Fortuny, and M. Stankova. Data mining for fraud detection using invoicing data. A case study in fiscal residence fraud. *University of Antwerp, Working Papers*, 2013.

[15] D. Martens, F. Provost, J. Clark, and E. Junqué de Fortuny. Mining fine-grained consumer payment data to improve targeted marketing. Technical report, Stern School of Business, New York University, 2013.

[16] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. *AAAI Workshop on Learning for Text Categorization*, 1998.

[17] W. Verbeke, K.. Dejaeger, D. Martens, J. Hur, and B. Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211–229, April 2012.