Mach Learn (2018) 107:1013–1037
https://doi.org/10.1007/s10994-018-5699-z

Wallenius Bayes

Enric Junqué de Fortuny[1] · **David Martens**[2] · **Foster Provost**[3]

Received: 29 October 2014 / Accepted: 15 February 2018 / Published online: 22 February 2018

Abstract This paper introduces a new event model appropriate for classifying (binary) data generated by a "destructive choice" process, such as certain human behavior. In such a process, making a choice removes that choice from future consideration yet does not influence the relative probability of other choices in the choice set. The proposed Wallenius event model is based on a somewhat forgotten non-central hypergeometric distribution introduced by Wallenius (Biased sampling: the non-central hypergeometric probability distribution. Ph.D. thesis, Stanford University, 1963). We discuss its relationship with models of how human choice behavior is generated, highlighting a key (simple) mathematical property. We use this background to describe specifically why traditional multivariate Bernoulli naive Bayes and multinomial naive Bayes each are suboptimal for such data. We then present an implementation of naive Bayes based on the Wallenius event model, and show experimentally that for data where we would expect the features to be generated via destructive choice behavior Wallenius Bayes indeed outperforms the traditional versions of naive Bayes for prediction based on these features. Furthermore, we also show that it is competitive with non-naive methods (in particular, support-vector machines). In contrast, we also show that Wallenius Bayes underperforms when the data generating process is not based on destructive choice.

Editor: Bernhard Pfahringer.

✉ Enric Junqué de Fortuny
enric.junquedefortuny@nyu.edu

David Martens
david.martens@uantwerp.be

Foster Provost
fprovost@stern.nyu.edu

[1] NYU Shanghai, 1555 Century Ave, Shanghai, China

[2] Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium

[3] Information, Operations and Management Sciences, Stern School of Business, New York University, New York City, USA

Keywords Naive Bayes · Wallenius distribution · Destructive choice

1 Introduction

As humans interact more with technology, we are increasingly seeing data sets incorporating people's fine-grained behaviors. In this paper we present a new classification technique based on naive Bayes, which can take into account two properties of many such datasets: the *proportionality* of choice and its *destructiveness*.

Proportionality of choice is a concept borrowed from psychological research: we select choices proportional to our preference for them (Luce 1959). Destructiveness is present when a person's choice to take a particular action removes that action from that person's future consideration. For example, you might "Like" Sheldon Cooper on Facebook at most once. After you do, Sheldon will not be in your subsequent consideration set. Similarly, you might not want to watch the same movie or read the same book twice.

Despite its simplicity, naive Bayes (Langley 1992; Domingos and Pazzani 1997) is a workhorse of machine learning in practice. Its assumption of class-conditional independence between feature probabilities renders the model simple to understand and to implement. Since it also exhibits relatively low variance (Ng and Jordan 2002), it is particularly accommodating for high-dimensional, sparse feature data, such as often are exhibited in data based on fine-grained human behavior (Junqué de Fortuny et al. 2013). Naive Bayes is based on an underlying event model, the dominant alternatives being multivariate Bernoulli (MV) and multinomial (MN) (McCallum and Nigam 1998). Following common practice, let us call the corresponding methods for learning and inference Multivariate Bernoulli Naive Bayes (MVNB) and Multinomial Naive Bayes (MNNB).

This paper addresses the concern that neither the MV nor the MN event model is a good approximation for destructive choice behavior, and therefore they might be expected not to perform well for prediction problems where the data are generated by such a process. The proposed *Wallenius* event model is based on the somewhat forgotten non-central hypergeometric distribution introduced by Wallenius (1963). We discuss its relationship with models of how human choice behavior is generated, highlighting a key (simple) mathematical property. Our main claim is that the Wallenius event model produces a version of naive Bayes (Wallenius Bayes) that is more appropriate for building predictive models from destructive behavioral choice data.

In the coming sections we will first frame the motivating principles behind the proposed method. We will illustrate these with an example and show how they match well to the Wallenius distribution. To support our main claim, we experiment with the three versions of naive Bayes on data where we would expect the features to be generated via destructive choice behavior. While not the main focus of the paper, we also compare Wallenius to the Support Vector Machine (SVM), which often performs very well on classification tasks. We close the paper by pointing out that although Wallenius Bayes is "naive" in the usual sense of naive Bayes, the event model creates a dependency on the order of choices in the true data-generating process. This matches the behavioral theory we present next, but also renders the implementation of Wallenius Bayes more computationally intensive than the alternative naive Bayes methods, opening an interesting avenue for future work.

2 Properties of human choice behavior

We will assume the standard supervised learning setting, where for a set of instances a certain set of features of these instances will be used to estimate or predict the value of some other, unknown but valuable, variable about the instance. We will presume that our instances are people, the features indicate actions they have chosen to take, and the target variable is some other trait of interest. We will not concern ourselves with theoretical justifications for why certain behaviors ought to predict other traits.[1] What is important here is that choice processes may place constraints on the sort of behavior data that may be generated.

2.1 Destructive choice

As introduced above, the important aspect of human behavior on which this paper focuses is the destructive nature of certain choices. Various sorts of constraints produce destructive choice. Often, after having made a choice, duplicate choices simply do not make sense. This is the case in our Facebook Likes example (and further, there the computer system embodies the constraint), as well as in the selection of keywords for research papers, the choice of players for one's fantasy football team, the phone numbers one puts in one's contact list, and many other settings.

Other problems are well approximated as destructive choice problems because of the constraints placed by what may be called "behavioral capital": a limitation on the time, money, and other resources a person can expend on the behaviors in question. For example, someone's partner may refuse to rent a movie for a second time. Thus, even though she might be convinced once in a blue moon to do so, her movie-viewing choice process would be well approximated as one of destructive choice.

2.2 Luce's choice axiom

There has been prior research focused on measuring choice behavior, and it is useful to ask whether our understanding of machine learning over choice behavior data can be informed or motivated by this prior work. Thurstone was one of the first to suggest a mathematical scale of psychological affinity with particular choices, proposing a "[discriminal process] where a single observer compares a series of stimuli by the method of paired comparison when no equal judgments are allowed" (Thurstone 1927). This scale implied the existence of a ranking of preferences and as such probabilities of a subject choosing one option over another. Others later noted that, although very often choice is a function of many influences and thus such a simple scale might not suffice, "it may be well to limit our scope to choice that exhibits both transitivity and unidimensional control [until we have succeeded in this task]" (Fantino and Navarick 1975). Transitivity enforces a consistent ranking, whereas unidimensional control captures our inability to account for external factors outside of the measured process. Experiments on pigeons followed to assess the validity of Thurstone's seminal work (Thurstone 1927). These experiments revealed that choice responses by pigeons indeed approximately match the proportions of reinforcements given (Herrnstein 1961).

[1] The interested reader may consider various theories of "inner drivers", including personal maxims (Kant 1790), moral inclinations (Kant 1790), social drivers (Bourdieu 1984) or simply a combination of experience and predisposition (Sapolsky and Bonetta 1997). All of these theories imply at least to some extent that there are correlations between our behaviors and our inner drivers, and that the realization of all of these aspects is part of what defines us as an individual.

For humans, it was Luce (among others), who produced the widely used generalization of this proportionality, called the *choice axiom*, leading to a definition of measurable choice probabilities (Luce 1959):

$$P_R(x_i) = w(x_i) \Big/ \sum_{j \in R} w(x_j) \ . \tag{1}$$

Here, $w(x_i)$ is the response strength (or the *preference*), associated with a response x_i. The choice axiom instantiates the proportionality via division by the total response strength over the choice set R.

In the problem setting for the present paper, we might envision a data generating process as follows. Given subjects of two particular groups (*classes*; e.g., smokers vs. non-smokers), each of the subjects makes various choices, one by one, based on her intrinsic drivers and preferences. These choices are destructive, in that a subject can choose an option only once.[2] Then in the standard mode of generative modeling, given a history of such choices can we predict what trait class a subject belongs to? The Wallenius Bayes model will tie the learning algorithm over the behaviors to the proportional choice probabilities of the Choice Axiom (Eq. 1), taking into account destructive choice behavior.

2.3 The learning task and notation

Formally, we model the choices of an individual i as a vector $\boldsymbol{x}_i$ and as mentioned above adopt the standard supervised classification formulation in which we want to predict labels y_i (e.g., gender) based on the choices individual i has made. Each of the individual scalar entries $x_{i,j}$ in $\boldsymbol{x}_i$ is Boolean valued and represents whether user i has chosen an item j or not. The labeled data then comprise a set of n examples $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where the label y_i represents the class of the individual (encoded as either 0 or 1). The question we concern ourselves with is whether we can predict the class memberships for previously unseen individuals. Can we find a function $f : \mathcal{X} \rightarrow \{0, 1\}$ such that $f(\boldsymbol{x}_i)$ predicts y_i well for any individual i making choices in the known feature set?

2.4 The data generating process: an illustrative example

Imagine going to the movies and finding yourself faced with a choice of m movies. If want to model the probability that you will choose any particular movie, we need to know and quantify your personal preferences. That is, you are likely to prefer some movies over others and this can be encoded with a preference weight w_j for any particular movie j. According to the reasoning presented in Sect. 2, the chances of you picking any particular movie are proportional to its preference score w_j and inversely proportional to the total preference weight of the movies presented at the movie theater.[3]

It is clear that you may go to the movies more than once so let us for this example assume that you go to the movies k times. Complicating things further, we might also want to account for the fact that a movie is not always screened at your local theater. Perhaps the next time you visit the theater, they will screen a new movie instead of an older one. Generally speaking, we say that each of the movies is presented to you as an option m_j times. Figure 1 shows

[2] Furthermore, constraints on behavioral capital will limit the total number of choices she can make. This is interesting in that it can explain the extreme sparsity of many behavioral data sets (Junqué de Fortuny et al. 2013), and can lead to efficient algorithm implementations, but is not essential for the present theory.

[3] While the notation differs slightly, these w_j directly map to the $w(x_j)$ weights in the choice axiom (Eq. 1) and can be considered to be the same.

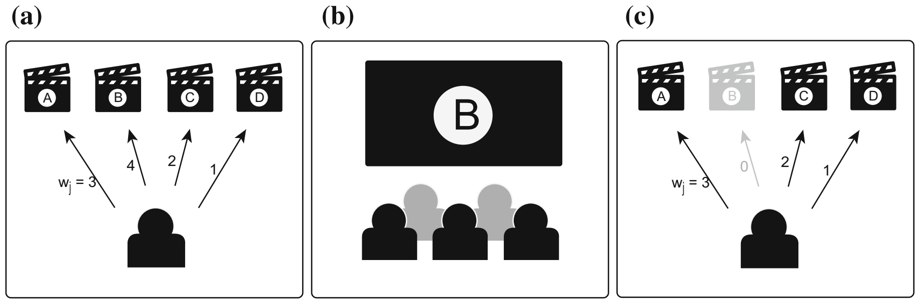

Fig. 1 Illustration of how a destructive choice process occurs in movie selection. **a** In the beginning of the process, the viewer has a fixed preference or utility (w) for each movie. **b** Based on this preference, she chooses to watch the second movie. **c** Having watched this movie, it is now very unlikely that she will watch it again and thus her preference for movie B needs to be updated accordingly

how the concept of *destructive choice* then applies to this scenario: after picking a movie (movie B in this example), your preferences change dynamically. As (in our model) you never want to watch a movie twice, subsequently you have a preference score of zero for that movie—movie B is removed from your choice set.

More generally, the probability of watching a set of movies $\boldsymbol{x}$ with x_j personal "views" for movie j can be represented by a multivariate Wallenius' non-central hypergeometric distribution with parameter vectors: movie history $\boldsymbol{x}$, number of screenings presented $\boldsymbol{m}$ and preferences $\boldsymbol{w}$. We will elaborate on this distribution in the following sections, but let us for now first calculate the movie history probability for our toy example to capture the intuition behind the distribution.

Consider a scenario containing four movies $\{A, B, C, D\}$. For this example we will assume movie preferences for a person being $w_A = 3$, $w_B = 4$, $w_C = 2$, $w_D = 1$. What then would be the probability of a particular movie-watching history, say B then A then C?[4] This particular history would have a probability of:

$$\begin{aligned} P(B \rightarrow A \rightarrow C) &= P(B) \cdot P(A \mid B) \cdot P(C \mid A, B) \\ &= \frac{4}{10} \cdot \frac{3}{6} \cdot \frac{2}{3} \\ &= 13.3\% \end{aligned} \tag{2}$$

Note how the denominator changes from ten ($\frac{4}{10}$) to six ($\frac{3}{6}$) and from six to three ($\frac{2}{3}$) after removing choices B and A respectively from the set.

Now, as with many data that are collected on human actions, let us assume that we do not actually observe the order, but just which movies were picked by each individual. This makes it more difficult to calculate the probability of the observation: we would need to sum the probabilities of all the possible orderings of movie views. We can represent the set of all possible movie view orders using a permutation set $\sigma(\boldsymbol{x})$. We denote one such ordering as an ordered set $\mathcal{X} \in \sigma(\boldsymbol{x})$. For instance, in the above example $B \rightarrow A \rightarrow C$ is one possible instantiation of $\mathcal{X}$. Another could be $A \rightarrow C \rightarrow B$, and so forth. In total, there are $k!$ such instantiations or permutations for a selection of k specific movies from a universe which contains M movies.

[4] Note that we restricted the number of seen movies in the history of this example to three for brevity and clarity reasons. We could of course expand this logic to longer sequences as well.

It is clear that each of the $\mathcal{X}$ belonging to permutation set contains the same elements, but in a different order. As such, we could also say that each ordering $\mathcal{X}$ indexes (~) the sequence differently:

$$\begin{aligned}
&\mathcal{X}_1 \sim x_{1,1} \rightarrow x_{1,2} \rightarrow \ldots \rightarrow x_{1,k}\\
&\mathcal{X}_2 \sim x_{2,1} \rightarrow x_{2,2} \rightarrow \ldots \rightarrow x_{2,k}\\
&\quad\vdots\\
&\mathcal{X}_j \sim x_{k!,1} \rightarrow x_{k!,2} \rightarrow \ldots \rightarrow x_{k!,k}.
\end{aligned}$$

Averaging the probabilities of all of the above sequences through summation and normalization over the total number of possible permutations in the population, then reveals the estimated probability of having watched this particular set of movies (without observed order):

$$\begin{aligned}
P(\boldsymbol{x}) &= P(\{A, B, C\})\\
&= \frac{1}{k!} \sum_{\mathcal{X} \in \sigma(\boldsymbol{x})} P(\mathcal{X}) \qquad (3)
\end{aligned}$$

In our example from before, Eq. (3) evaluates to:

$$\begin{aligned}
P(\boldsymbol{x}) = \frac{1}{3!} \cdot \Big[&P(A \rightarrow B \rightarrow C) + P(A \rightarrow C \rightarrow B) + P(B \rightarrow A \rightarrow C)\\
&+ P(B \rightarrow C \rightarrow A) + P(C \rightarrow A \rightarrow B) + P(C \rightarrow B \rightarrow A) \Big]\\
= \frac{1}{6} \cdot &\left[\frac{24}{210} + \frac{24}{350} + \frac{24}{240} + \frac{24}{180} + \frac{24}{400} + \frac{24}{320} \right]\\
= 9.19\%&
\end{aligned}$$

Note again how the denominator changes to reflect the destructive choice being taken into account for different orderings. Generalizing for sequences of arbitrary size k, we get that:

$$\begin{aligned}
P(\boldsymbol{x}) &= \frac{1}{k!} \sum_{\mathcal{X}_j \in \sigma(\boldsymbol{x})} P(x_{j,0} \rightarrow x_{j,1} \rightarrow \ldots \rightarrow x_{j,k}) && (4)\\
&= \frac{1}{k!} \sum_{\mathcal{X}_j \in \sigma(\boldsymbol{x})} \prod_{t=1}^{k} P(x_{j,t} \mid x_{j,t-1} \ldots x_{j,1}) && (5)\\
&= \frac{1}{k!} \sum_{\mathcal{X}_j \in \sigma(\boldsymbol{x})} \prod_{t=1}^{k} \frac{w_{j_t}}{\sum_{t' \geq t} w_{j_{t'}}} && (6)\\
&= \frac{1}{k!} \sum_{\mathcal{X}_j \in \sigma(\boldsymbol{x})} \prod_{t=1}^{k} \frac{w_j}{\sum_{j=1}^{M} w_j - \sum_{t' < t} w_{t'}} && (7)
\end{aligned}$$

In Eq. (5), we combine two elements from before: (a) a multiplication over all elements in a sequence of draws in "time" (position) t to get the probability of that particular sequence (Eq. 2) and (b) a sum over all possible sequence permutations $\sigma(\boldsymbol{x})$ (Eq. 3). As before, the normalizing constant $k!$ is equal to the total number of possible orderings for the choices—the orders one could have watched k movies from a set of M movies. In Eq. (6) we then write the fractions explicitly by dividing the weight w_{j_t} of watching movie j_t (chosen at time t), by the weights of the choices (movies) that are still left in the theater at that choice point in the sequence. This is then expanded to arrive at the final formulation in Eq. (7).

Note the close resemblance of Eq. (6) with the proportionality of response strengths given by the choice axiom when dealing with human choice behavior (Eq. 1). The right side of Eq. (6) describes a process that follows the choice axiom, but here we look at the joint probability of all events that were produced by such a process (product over t) and average over sequences since we do not observe the actual sequence (sum over $\sigma(\boldsymbol{x})$). It is then easy to see that the movie history probability is indeed in line with the probabilities resulting from human behavior given a sequence of historical choices (we show this in more detail in the Appendix).

Unfortunately the direct computations can take a long time. Using an efficient depth-first tree method with memory still requires $\mathcal{O}(k!)$ computations, because we need to compute each of the permutations individually. In Sect. 4.1, we will introduce a more efficient approximation of the above distribution, after briefly discussing the alternative naive Bayes event models.

3 Naive Bayes event models

In this section, we describe how traditional event generating distributions used with Naive Bayes would model the data presented in the previous section, and discuss the associated limitations for the destructive choice setting.

3.1 Multivariate Bernoulli Naive Bayes

The multivariate Bernoulli event model for naive Bayes (MVNB) assumes that the data are generated according to an independent Bernoulli process for each of the M features. The probability of an observed vector $\boldsymbol{x}$ (with feature values x_j for each feature X_j), given that the underlying observations result from the behavior of a particular class y, is then simply the product of all of the independent Bernoulli probabilities $P(X_j = x_j \mid Y = y)$. The resulting class-conditional probability then becomes:

$$P(\boldsymbol{x}_i \mid Y = 0) = \prod_{j=1}^{M} \left[P(X_j = x_j \mid Y = 0)\right]^{x_{i,j}} \left[1 - P(X_j = x_j \mid Y = 0)\right]^{(1-x_{i,j})}, \quad (8)$$

where the powers $x_{i,j}$ serve as a selection mechanism since they are either 1 or 0, depending on the observed value. Optimizing the log likelihood, equating to zero and taking a Laplacian prior, yields the following parameter estimates for binary data (McCallum and Nigam 1998):

$$\hat{P}(X_j = x_j \mid Y = 0) = \frac{1+ \mid X_j = 1 \wedge Y = 0 \mid}{2+ \mid Y = 0 \mid}, \quad (9)$$

where we used the typical bar notation $\mid \mathcal{C} \mid$ to refer to the count of elements in the set that satisfy condition $\mathcal{C}$.

Naive Bayes based on the multivariate event model is very fast to compute, but by design does not take into account the influence of context or correlations of the choices. 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 (Flach and Lachiche 2000, 2004). In MVNB, each choice is treated as a coin-flip with a fixed probability, estimated on the training data. As shown in the

Appendix this leads to a violation of the choice axiom, and thus renders the model dubious for faithfully representing choice behavior based on preference. Furthermore, MVNB does not take into account the destructive nature of the choice data (see Sect. 4.2). The result is that MVNB produces highly skewed posterior probabilities for human behavioral data sets.

3.2 Multinomial Naive Bayes

The main idea behind the multinomial event model is that each input sample is the result of a series of independent draws from a bag of items (the choices, corresponding to the features). To determine the items present in an input vector $\boldsymbol{x}_i$, a total of k_i independent draws is undertaken from an underlying multinomial distribution, each time picking one item from all the M possible items $(X_1 \dots X_M)$ with replacement. The probability of one such draw (picking a feature X_j from all possible features) is modeled as $P(X_j = x_j \mid Y = y)$, dependent only on the count evidence seen in the historical data. The aggregated probability of seeing the input vector given that it belongs to a certain class can then be modeled as a multinomial:

$$P(\boldsymbol{x}_i \mid Y = 0) = (k_i!) \cdot \prod_{j=1}^{M} \frac{P(X_j = x_{i,j} \mid Y = 0)^{x_{i,j}}}{x_{i,j}!}. \tag{10}$$

The probability parameters $\theta_{X_j = x_j | Y = y} = P(X_j = x_{i,j} | Y = y_y)$ for this model are the probabilities for each feature occurring, conditional on the class (i.e., these are estimated for each class separately). Assuming a Laplacian prior $\hat{\theta}_c$ on the parameters, we obtain the Bayes-optimal (maximal likelihood) estimates of the parameters for the formulation in Eq. (10) (McCallum and Nigam 1998):

$$P(X_j = x_j \mid Y = 0; \hat{\theta}_c) = \frac{1 + \sum_{i=1}^{n} x_{i,j} P(Y = 0 \mid \boldsymbol{x}_i)}{M + \sum_{j=1}^{M} \sum_{i=1}^{n} x_{i,j} P(Y = 0 \mid \boldsymbol{x}_i)}.$$

In the case of binary data this becomes:

$$\hat{\theta}_{X_j = x_j | Y = 0} = \frac{1 + \mid X_j = 1 \wedge Y = 0 \mid}{M + \sum_{k=1}^{M} \mid X_k = 1 \wedge Y = 0 \mid}, \tag{11}$$

where we used the $\mid \mathcal{C} \mid$ notation again to indicate the count of elements over the full data set that satisfy condition $\mathcal{C}$.

One of the main advantages of this formulation is that it only requires two passes over the non-zero ("active") elements for all of the parameters to be estimated. As we show in the Appendix, MNNB does not violate the choice axiom and might thus seem to be a better alternative than MVNB for the choice data we are considering. Unfortunately, MNNB is not well aligned with our problem setting either. The reason is that as it draws with replacement from the bag of items, it does not take into account destructive choice. This means that if one were repeatedly to visit a movie theater, all previous movie choices would be ignored (i.e. each visit would look the same a priori). As we will see in Sect. 4.2, this also leads MNNB to produce skewed posterior probability estimates.

4 Wallenius Bayes

In this section we first describe the general form of the Wallenius distribution which lies at the core of the new event model, followed by some refinements and technical remarks on implementation.

4.1 Wallenius's non-central hypergeometric

4.1.1 General form

Comparing to the brute force representation of destructive choice presented in Sect. 2.1, a more elegant solution to the ordering problem mentioned in Sect. 2.4. was found by Wallenius (1963) [and later generalized in Chesson (1976)] 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:

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

As before, $\boldsymbol{m}$ is a vector representation of the number of times each choice can be made—in our running example, the number of screenings of each movie—and $\boldsymbol{w}$ a vector of the (preference) weights corresponding to the choices. Note that in our example, $\boldsymbol{x}$ is a binary vector of length M, containing a one ($x_j = 1$) if movie j was watched and a zero ($x_j = 0$) otherwise. Note that this expansion does not explicitly consider ordering information; this is one of its main strengths since it avoids the explicit summation over all of the permutations.

4.1.2 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 ($\forall j : x_j \in \{0, 1\}$) and assuming a uniform one-shot distribution of presented choices ($\boldsymbol{m}$ is a vector of ones), we can rewrite the Wallenius distribution as:

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

This simplifies the computation substantially. Note also how no equation is dependent on **m** anymore; we will make the uniform screening assumption explicit by dropping $\boldsymbol{m}$ as an input to the distribution when dealing with the binary case. Even so, the fractional exponent in the simplified form still makes the integral difficult to compute efficiently using typical numerical integration methods. Fortunately, we can simplify the integrand to a polynomial using variable substitution ($u = \sqrt[s]{t}$):

$$wall(\boldsymbol{x}; \boldsymbol{w}) = s \cdot \int_0^1 u^{s-1} \cdot \prod_{j|x_j=1} \left(1 - u^{w_j}\right) du. \tag{12}$$

This accounts for all non-degenerate cases, the exception being the degenerate case of $s = 0$ (having watched all the movies) which should of course return 1.

Similar to other event models, there is a degeneration of the accuracy when features are completely unobserved in the training data: $w_j = 0$ for any $j \mid x_j = 1$ causes $wall(\boldsymbol{x}; \boldsymbol{w}) = 0$, which is not the desired behavior. We can remedy this in the usual manner by adding a *pseudo-count* to the weight vectors that determine how telling each movie is for the class prediction (i.e., each movie's weight is set to one before model training begins). This corresponds to the prior belief that every feature is equally probable and results in a uniform smoothing of the probabilities that decreases as the evidence grows.

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

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

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

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:

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

When fewer steps are used, the approximation suffers from an accuracy degeneration bounded by (Kahaner et al. 1989):

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

with I the integrand with $2n = order$ continuous derivatives. 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 paper [and a comprehensive overview is given in Fog (2008)]. In our experiments we could deal with up to a thousand instances using thousands of features in a reasonable amount of time on a low-end computer. Scaling up further is an avenue for future work.[5]

[5] An open source implementation is available on http://github.com/ciri.

4.1.3 Using the model for prediction

As explained in the previous section, the final form given in Eq. (12) can be computed relatively easily. To actually predict the class membership of a test instance, one would have to calculate the class conditional probabilities for each class separately and compare them. One easy way to aggregate information on both classes is to look at the normalized difference between both conditional probabilities:

$$P(Y = y_i \mid \boldsymbol{x}) \propto P(Y = y_i) \cdot P(\boldsymbol{x} \mid Y = y_i) \tag{13}$$

$$\begin{aligned} Score(\boldsymbol{x}) &= \frac{P(Y = 1 \mid \boldsymbol{x}) - P(Y = 0 \mid \boldsymbol{x})}{P(Y = 1 \mid \boldsymbol{x}) + P(Y = 0 \mid \boldsymbol{x})} \\ &= \frac{P(Y = 1) \cdot P(\boldsymbol{x} \mid Y = 1) - P(Y = 0) \cdot P(\boldsymbol{x} \mid Y = 0)}{P(Y = 1) \cdot P(\boldsymbol{x} \mid Y = 1) + P(Y = 0) \cdot P(\boldsymbol{x} \mid Y = 0)}. \end{aligned} \tag{14}$$

The sign of the score indicates the predicted class and estimation certainty is represented by large absolute values of this score, representing starkly different predicted class membership probabilities in the comparison.

In order to calculate the score, we applied the Bayes' rule and the common (but not necessary) assumption that each input sample is as likely to occur as any other [revealing the likelihood in Eq. (13)]. The maximum likelihood (point) estimate for the prior class probability $P(Y = y_i)$ can be calculated by looking at the fraction of samples belonging to class C_i in the total training set. The class conditional probabilities $P(\boldsymbol{x} \mid Y = y_i)$ can be calculated using $wall(\boldsymbol{x}; \boldsymbol{w}_i)$, where $\boldsymbol{w}_i$ is a weight vector for class i. Weight vector $\boldsymbol{w}$ is estimated from the training set and requires a linear pass through the active (non-zero) elements of the data. That is, we can calculate $\boldsymbol{w}$ by summing the number of movie watches of people from the relevant class (this is simply a conditional column sum over the matrix if we represent the subjects as rows and behaviors as columns). Making predictions requires two computations of the Wallenius class-conditional probability estimate (one for each class).

4.2 Comparison with other event models

We mentioned before that using the multivariate Bernoulli event model or the multinomial event model can result in skewed posterior probabilities. In this section we present an elaboration of our toy example to illustrate the effects of the destructive choice setting, where the Wallenius event model would match much better to our intuition than either of these traditional event models.

Let us consider a scenario in which we try to predict gender based on movie-viewing history in a movie theater where they screen only 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).

Table 1 Artificial data set of movie-watching behavior of 100 male (M) and 100 female (F) subjects

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

Note that the sum of a row might be larger than 100 because any single subject may watch more than one movie

4.2.1 Difference 1: What happens when additional choices are presented?

Consider if the data were extended to reveal an additional choice (movie) with a high preference weight. A subject would be inclined to make that choice at least to some extent and this should be accounted for in a model's individual probabilities. Let us see how each model deals with the inclusion of an additional movie in our example, all else being equal (a more formal treatment is given in the Appendix).

In the initial state displayed in Table 1, the probabilities for movie 1 (M_1) according to the multivariate Bernoulli (MV) and the Multinomial (MN) model are respectively:

$$P_{MV}(M_1 \mid \text{Male}) = \frac{90}{100} = 90\%$$
$$P_{MN}(M_1 \mid \text{Male}) = \frac{90}{110} = 81.82\%.$$

Due to the complete independence of other choices, the Bernoulli probabilities would be unchanged after adding the new choice (each of the $P(M_i \mid M)$ is constant). In contrast, the multinomial model (and the Wallenius model) would take the context into account via the renormalization of the probabilities. As such, observing say a fourth blockbuster movie with 100 watches for each gender, would result in:

$$P_{MV'}(M_1 \mid \text{Male}) = \frac{90}{100} = 90\%$$
$$P_{MN'}(M_1 \mid \text{Male}) = \frac{90}{110+100} = 40.91\%.$$

Indeed, if our favorite movie theater were to add a blockbuster movie that we are very eager to see, that might alter our initial choice of movie even if we did not have the blockbuster in mind initially. Note, however, that in accord with the choice axiom, the relative proportionalities of the probabilities (their ratios with respect to one another) do not change for MNNB.

4.2.2 Difference 2: Dealing with destructive choice

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). She decides to see M_3. Can we predict Sophie's gender given only her movie watching history?

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

$$P(M_1, M_3 \mid \text{Male}) = \frac{90}{100} \cdot \frac{10}{100} \cdot \left(1 - \frac{10}{100}\right) = 8.10\%$$
$$P(M_1, M_3 \mid \text{Female}) = \frac{90}{100} \cdot \frac{10}{100} \cdot \left(1 - \frac{1}{100}\right) = 8.91\%.$$

Due to the independence assumption, both genders are almost equally likely with a resulting score (recall Eq. 14) of $S_B(M_1, M_3) = 0.0476$ (slightly in favor of female).

Table 2 Summary of differences between various underlying event models for naive Bayes

	Choice changes preferences	Proportionality of preferences constant
Bernoulli	No	No
Multinomial	No	Yes
Wallenius	Yes	Yes

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

$$P(M_1, M_3 \mid \text{Male}) = \frac{90}{110} \cdot \frac{10}{110} = 7.44\%$$
$$P(M_1, M_3 \mid \text{Female}) = \frac{90}{101} \cdot \frac{10}{101} = 8.82\%.$$

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 with equal probability.

Of course, ideally in our example, the probability of seeing a previously seen movie should plummet to near zero. Thus, Sophie's choosing the balanced-gender movie rather than the male-dominated movie ought to be a telling event. However, since the multinomial event model keeps all three movies in consideration, the result is only a relatively small increase in the estimated probability that she is female (resultant score $S_M(M_1, M_3) = 0.0849$).

Wallenius's answer The Wallenius event model considers the data to have been generated via a destructive choice process. However, we assume that we have the same data as in the traditional setting—namely, that we do not know the actual sequence of events. Thus, the Wallenius model for the conditional probability averages over all possible sequences:

$$P(M_1, M_3 \mid \text{Male}) = \frac{1}{2} \cdot \left(\frac{90}{110} \cdot \frac{10}{110 - 90} + \frac{10}{110} \cdot \frac{90}{110 - 10} \right) = 24.55\%$$
$$P(M_1, M_3 \mid \text{Female}) = \frac{1}{2} \cdot \left(\frac{90}{101} \cdot \frac{10}{101 - 90} + \frac{10}{101} \cdot \frac{90}{101 - 10} \right) = 45.40\%.$$

Although there is still some bias due to the fact that we do not know the true sequence order, the conditional probabilities reflect more certainty, resulting in a score of $S_W(M_1, M_3) = 0.298$.[6]

The conclusions of this section are summarized in Table 2, which should also give some insights as to when to use which method.

5 Empirical evaluation

The example discussed in the previous section was constructed artificially and it stands to reason that the models might behave differently when applied to destructive-choice data from empirical settings. In this section, we compare the performance of all of the previously mentioned methods, using human behavioral data from various settings where destructive

[6] Had we actually known the true ordering, it would have been easier to discriminate between both with the probabilities being 40.91% for the male and 81.01% for the female case (leading to a score of 0.329).

Table 3 Overview of destructive choice data sets

Dataset	n	d	Order	Prediction task
*Politic	8515	132,351	$\geq$ 1,975,761	Liberal versus democrat
*Religion	13,692	183,384	$\geq$ 2,783,392	Catholic versus Islam
†IQ	6377	134,420	$\geq$ 919,684	Subject has high IQ
*Gay M	2491	54,106	$\geq$ 433,561	Exclusive same-sex interest for men
†Smoking	3746	95,186	$\geq$ 821,457	Daily smoking behavior
Yahoo Movies	7642	11,917	$\geq$ 220,809	Gender of user based on movie views
Book-Crossing (Ziegler and McNee 2005)	118,270	61,309	$\geq$ 618,886	Age based on book interests
Ta-Feng	31,640	23,721	$\geq$ 723,449	Age based on product interest
Dating	135,359	220,970	$\geq$ 17,359,099	User is attractive based on social network

The data sets above the horizontal center line use Facebook Likes as choices and originate from (Kosinski et al. 2013). For the Facebook tasks, the values for the target variables are based either on information revealed by the user (indicated by a *) or tests (Kosinski et al. 2013; Etter et al. 2003) (indicated by †)

choice is a reasonable approximation. In doing so, we want to address the following four questions, presented in order of importance:

1. Is the Wallenius event model indeed more appropriate in settings where destructiveness and proportionality of choice influence the probability regime?
2. At which sample/data set size does the advantage of its richer model representation stop mattering?
3. Is Wallenius naive Bayes competitive with other state-of-the-art methods?
4. How well does Wallenius naive Bayes fare in settings outside of the proposed destructive/proportional choice context?

As noted before, the computational difficulty for each experimental setting depends on the maximal order of the integral. As shown in Table 3, this order varies between data sets and is not necessarily dependent on the number of choices or instances, but rather on the number of active elements (choices made). In essence, the sparser the data, the easier it is for Wallenius Bayes to come up with a solution fast. For the experiments conducted in this setting, we set the maximum number of steps in the integral approximation to 2000, leading to an acceptable speed-accuracy trade-off with most models being generated in a matter of seconds.

5.1 Is Wallenius a better alternative for destructive-choice data?

In order to know whether Wallenius is a better alternative we compare the performance of naive Bayes using the three alternative event models experimentally on empirical data.

Let us start by focusing on the Yahoo Movies data set. As shown in Table 3, this data set is a collection of 7, 642 users' movie choices. In total, these users watched a quite diverse selection of 11, 917 movies. The total number of choices recorded is 220, 809, or about 29 movies per person on average. This results in a very sparse dataset; this is typical for choice data (Junqué de Fortuny et al. 2013).

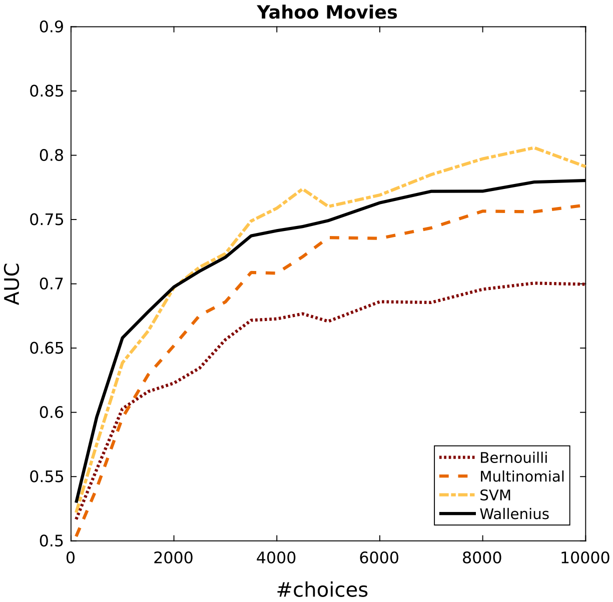

Fig. 2 Experimental results for MV Bernoulli, the multinomial and the Wallenius event models on the Yahoo Movies data set for increasing numbers of choices. The Wallenius event model outperforms the other methods in terms of AUC for increasing numbers of randomly selected choices. Also shown is the performance of linear support vector machines

Similar to the toy example, the goal is to predict the gender of the user, based on which movies the person has chosen.[7] As explained before, one would expect the data generating process to be both destructive and proportional. We therefore expect Wallenius to do better than the multinomial event model, which in turn should do better than the multivariate Bernoulli event model (due to proportionality being modeled).

The results of the experiment are shown in the Fig. 2. In the experiment, we assess predictive power on a held-out test set, evaluated in terms of Area Under the ROC curve (Fawcett 2006) which, equivalent to the Mann–Whitney–Wilcoxon statistic, measures how well a model's scores rank positive versus negative instances from a data set.

To assess the stability of the results and to later answer the second question, the experiments vary the choice sets used for each prediction task. Specifically, for each prediction task we experiment with randomly chosen choice sets of increasing size, i.e., the experiments incrementally build increasingly large choice sets and rerun the experiments on each. In the figures, the AUC values are plotted for these increasing numbers of choices. That is, the resulting curves resemble standard learning curves, but instead of increasing numbers of instances, the horizontal axis represents increasing numbers of choices.

Each point on each curve is the average AUC over ten training/testing splits. For each, the data set is randomly split into a training set containing 90% of the data and a testing set (held out during training), containing the remaining 10% of the data and used to estimate the generalization performance. The results therefore not only compare across prediction tasks but also across choice set sizes.

[7] Fating a movie is used as a proxy for a person having watched it.

The Yahoo curve is typical for the experiments to come and shows that the Wallenius event model is indeed better than the other event models for this data set. The black curve (Wallenius) dominates the orange dashed curve (the multinomial event model) which itself dominates the red dotted curve (the multivariate event model). These results therefore support our insights from Sect. 4.2.

We also evaluate the Wallenius Bayes event model on other prediction tasks that should be well approximated by destructive choice. Five of the prediction tasks are based on Facebook users' choices [shown in Table 3, first used in a study by Kosinski et al. (2013)]. For each of these data sets, the choice set comprises possible Facebook 'Likes': on Facebook, a user can choose to indicate whether or not she likes something. Importantly, a Facebook user indeed can only Like something once. We use these Likes to predict several of the users' personal traits. The personal traits either were revealed by the user (directly or indirectly) on Facebook or inferred from a psychological test. We also examine other data sets where predictions are made based on users' choices, shown below the horizontal center line in Table 3. These include data on movie watching (presented above), book reading, product purchasing, and friend choice.

We repeat the experiment described above; Fig. 3 compares the performance of naive Bayes with the three different event models for the eight remaining prediction tasks in a similar way as before, for varying numbers of choices. The results show superior performance for Wallenius Bayes over almost all of the data sets. Testing this result using a Wilcoxon signed rank test shows that this superiority is significant across the panel ($p < 0.02$). On the other hand, for these data there is no systematic superiority of the Multinomial event model over the Bernoulli event model.

Figure 5 (Appendix) studies the correlation between the different models using 1000 and 10,000 choices in the same experiment. We plot both the individual experiments (blue dots) as well as a diagonal line, corresponding to the theoretical equal performance line (i.e., dots on the line indicate identical performance for that experiment). The correlation coefficient between WNB and MVNB is higher ($r_{1000} = 0.72, r_{10,000} = 0.84$) than the one between WNB and MNNB ($r_{1000} = 0.36, r_{10,000} = 0.78$), indicating that the Wallenius and Multinomial event model perform similarly on the same problems. In the next section we investigate further how the multinomial and Wallenius event model differ. Generally, we see that most of WNB's results lie above the diagonal when comparing to both MVNB and MNNB, confirming the superior performance by Wallenius across different datasets.

These results confirm that taking into account the destructive choice setting can confer advantage when applying naive Bayes to human behavioral data.

Note that we see the strongest advantage for the Wallenius event model when modeling with smaller choice sets. This may be expected, as for larger choice sets the change in the renormalized probabilities after removing the "destroyed" choice is smaller, and so all else being equal the advantage of explicitly taking destructive choice into account ought to decrease. As the choice sets grow very large, the differences between Wallenius NB and MNNB or MVNB often decrease, with MNNB surpassing Wallenius NB in one case (Smoking) for the largest choice sets and with MBNB essentially indistinguishable from Wallenius NB for one data set (Dating).

5.2 Destructive choice versus alternatively generated data

Given the promising results from the previous section, it is important not to give the impression that Wallenius Bayes is simply a better naive Bayes. Therefore, let us examine how Wallenius

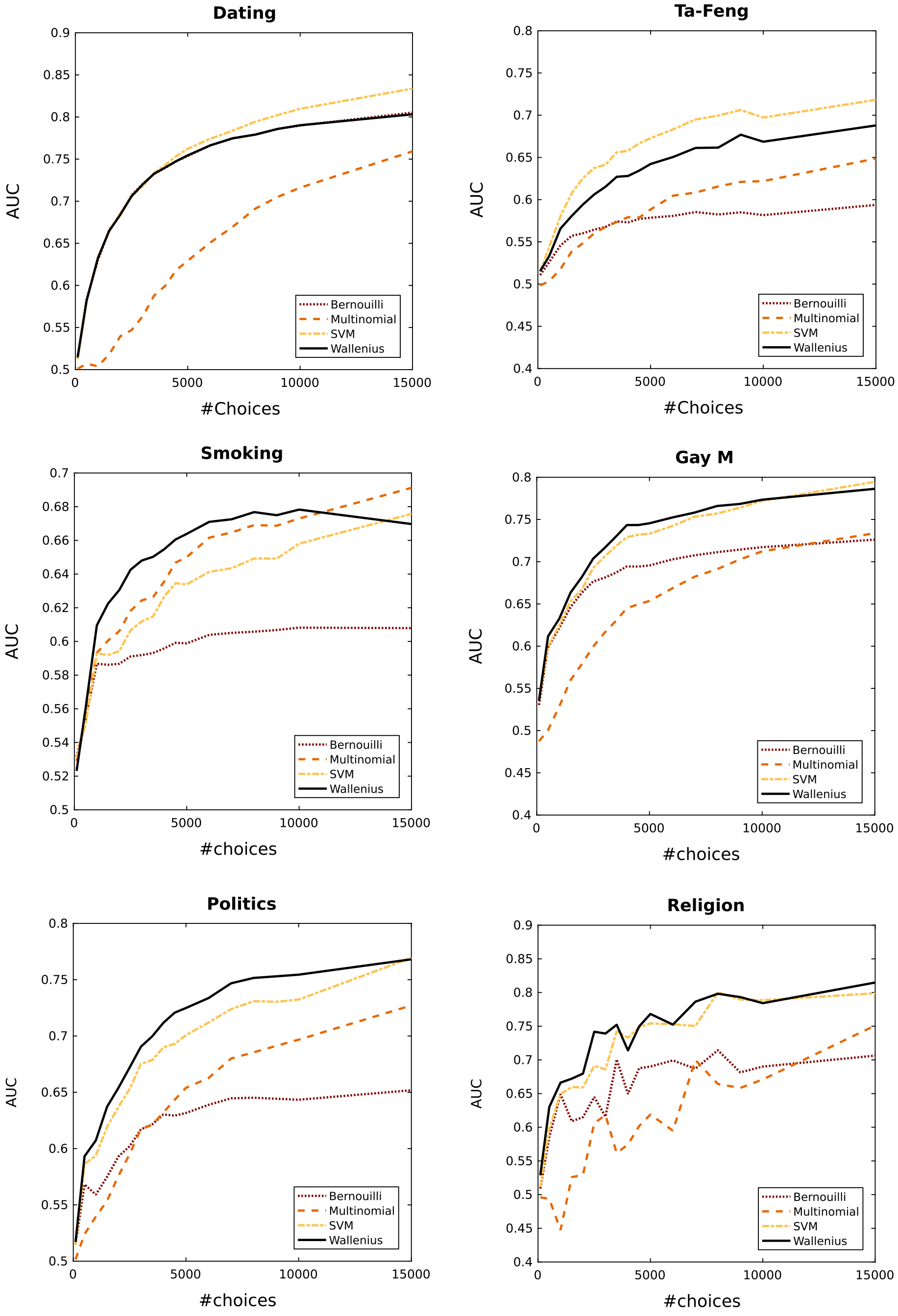

Fig. 3 Experimental results for the multivariate Bernoulli, multinomial and Wallenius event models on various data sets for increasing number of choices. In the majority of cases the Wallenius event model performs better than the other event models, in terms of AUC for increasing numbers of randomly selected choices. Also shown are curves for linear support vector machines

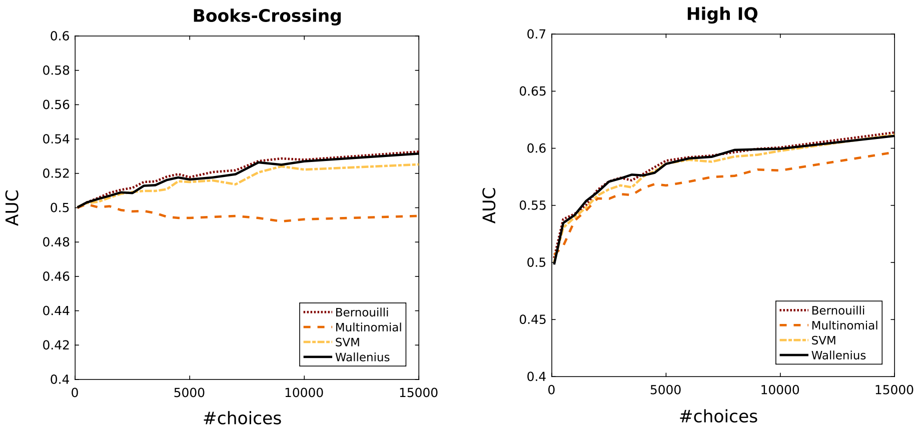

Fig. 3 continued

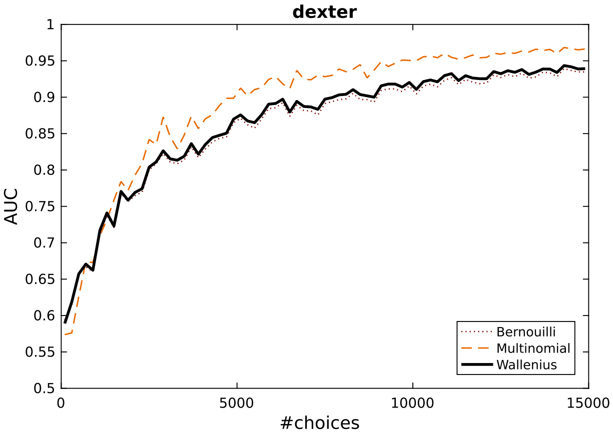

Fig. 4 Comparison of multivariate Bernoulli, multinomial, and Wallenius naive Bayes models for a text classification task. The data generating process for text documents is not a destructive choice process, and the multinomial event model clearly fares better than the destructive-choice-oriented Wallenius event model

Bayes responds to a data set where destructive choice would not characterize (or approximate) the data-generating process.

We compare the different versions of naive Bayes on the binarized Dexter subset of the famous RCV-1 data set (Asuncion and Newman 2007), where the prediction is whether the text is related to corporate acquisition or not. When writing documents, choosing a word does not exclude choosing that word again in the same document. It has been shown (McCallum and Nigam 1998) that multinomial naive Bayes works well for text classification, where each document is represented by either a vector of term frequencies or a binary value indicating whether a word is present in a document or not. Often MVNB is applied to binarized text-classification data, with success, even though doing so does not align well with the multivariate event model. When inspecting the learning curve in Fig. 4, we see that the outcome is as expected: the Wallenius and Bernoulli multivariate event models do not perform as well as the multinomial event model for this data set.

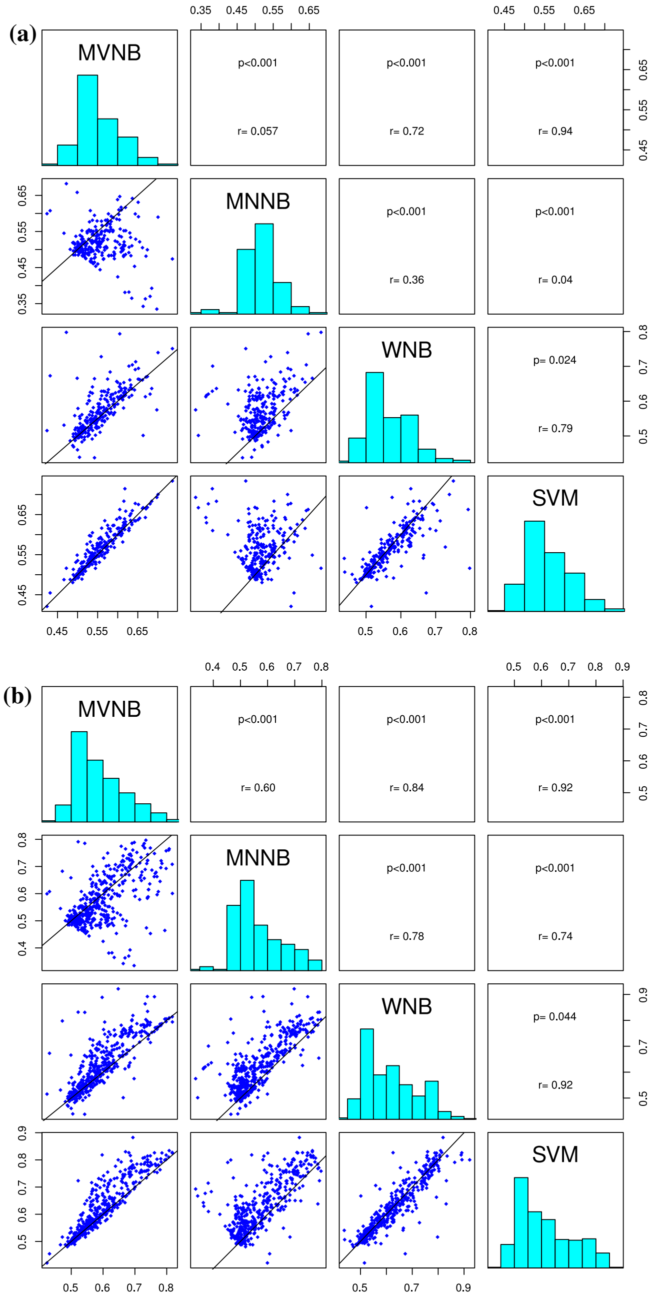

Fig. 5 Comparison of AUCs across models, over all experiments conducted in this study using 1000 and 10,000 choices. We plot both the individual experiments (blue dots) as well as the equal performance (diagonal) line. Compared to MVNB and MNNB, most of WNB's results lie above the diagonal, indicating superior performance by Wallenius. This superiority is significant with *p* values below 0.001 using a Wilcoxon signed rank test. Additionally, WNB compares comparably to (slightly better than) the SVM, with the correlation (r) becoming high as we consider a larger number of choices. **a** Results for 1000 choices. **b** Results for 10,000 choices (Color figure online)

5.3 Comparison with SVM

While not the main goal of this paper, to give additional context we also compare the three naive Bayes models on the behavioral data with a linear support vector machine (SVM). Figures 3 and 5 include curves (computed as described above) for each of the data sets.

When comparing with the SVM using the correlation analysis (Fig. 5), we notice that the SVM clearly does significantly better than MVNB and MNNB for both 1000 and 10,000 choices. Comparing versus Wallenius, we see that the models perform almost equally well. The test results indicate a (slightly) significant difference in model performance, with an advantage for Wallenius with 1000 choices and an advantage for SVM with 10,000 choices. While not shown here, similar results were found for lower and higher numbers of choices with the tipping point being at about $2,500$ choices. At the same time, we also notice a trend with both models becoming increasingly indistinguishable in data regimes with higher numbers of choices. This is also shown by an increase in correlation (r) in Fig. 5.

6 Discussion

Experimental results

From the above experiments, we conclude that for destructive choice data, the Wallenius event model appears to be generally superior to the multinomial and MV Bernoulli models. For small to intermediately sized data sets this conclusion extends to the SVM as well. In scenarios where we already have a lot of information or where the event model violates the prescribed conditions (proportionality, destructive choice) this advantage vanishes and other models may indeed become more appropriate.

Naiveness

The experimental results look very promising for the application of Wallenius Bayes in the context of destructive choice, which could be seen as a form of conditional dependence. This begs the question of whether we should still be calling it a “naive” method. We believe it should: the “naiveness” of Naive Bayes generally is considered to be the assumption of prior class-conditional independence of the choices. This assumption remains intact in Wallenius Naive Bayes—albeit the probabilities are renormalized in the context of considering the different possible orderings.

Ordering

Wallenius Bayes is slower than the alternatives (MVNB and MNNB). This is partly due to us not having information on the actual timing of events. If we had true sequences of events, we could simply apply Eq. (2), thus avoiding any expensive computations at all (this would essentially render the method linear in time complexity). As such, we hope that our results might be a precursor for more to come. To this end, research is needed into (a) the speeding up of Wallenius Bayes for non-sequence data and (b) the application of Wallenius Bayes to time-stamped data sets.

Ubiquity of the data generating process in online data

We have shown how the Wallenius method differs from more traditional event models theoretically (this was summarized in Table 2). While our method was motivated by explicit human behavioral choice data, it is not limited to such data. Indeed, as mentioned in Sect. 2.1 there are other scenarios in which the Wallenius method is applicable. First we have those scenarios where a system might embody the destructive constraint: Liking something on Facebook, favoriting a picture online, adding a product to one’s wish-list, adding a friend on

Twitter, the choice of players for one's fantasy football team, etc. Furthermore there are plenty of scenarios where the constraint is implicitly true: the selection of keywords for research papers, adding a phone number to one's contact list, etc. In terms of proportionality, there are domains for which it is well known that proportionality is appropriate. One such case is the proportionality of choice behavior in horizontally differentiable goods. It stands to reason that there will be many more scenarios that may be approximated using the proportionality of choice.

7 Conclusion

We presented a new event model for naive Bayes, based on the Wallenius distribution, and showed that it substantially outperforms the two traditional event models when working with behavioral data generated via destructive choice processes—which are common in the sort of data we increasingly are seeing as a result of monitoring human behavior. The model explicitly takes into account destructive choice and the proportionality of human preferences. The superiority of Wallenius Bayes may apply to other situations where destructive choice is part of the data generating process. We also show that in contexts where the data likely are not generated by a process well approximated by destructive choice—such as for text classification—Wallenius Bayes can be inferior to Naive Bayes with traditional event models.

There are several promising paths of future research that show potential. First, the performance of Wallenius Bayes for time-stamped data sets might show even greater advantage both in terms of accuracy as well as speed. For aggregated data (without timing information), further research into speeding up Wallenius Bayes is the obvious next step. A natural extension to the actual event model would be to include a length prior to the distribution as well using a more sophisticated Bayesian formulation. On a broader level, the work presented in this paper shows potential for rekindling the development of new event models to fit different scenarios as opposed to thinking of our algorithms simply as all-purpose learning methods. It may also be advantageous to consider whether considering destructive choice could be helpful beyond generative modeling.

Acknowledgements Thank you very much to Michal Kosinski, David Stillwell and Thore Graepel for sharing the Facebook Likes data set. Thanks to our reviewers for helpful feedback. David thanks the Flemish Research Council (FWO) for financial support (Grant G.0827.12N). Foster thanks NEC and Andre Meyer for Faculty Fellowships. We thank the Moore and Sloan Foundations for their generous support of the Moore-Sloan Data Science Environment at NYU.

Appendix: Event models and choice axiom

The choice axiom is stated in terms of the probabilities of choices. In this Section we will adopt the notation of Luce (1959) for sets of choices R, S, T, U, individual choices x, y and preference probability $P(x, y)$ (indicating how probable it is that x will be chosen over y in a pairwise comparison). A subindex $P_S(\cdot)$ indicates the preference probability in a set S. And set-probability $P(R)$ indicates the probability that any choice in R would be picked over the complement set $\overline{R}$. The relationship under investigation is then such that it follows Axiom 1, presented below.

Axiom 1 [The Choice Axiom; Luce (1959)] *Let T be a finite subset of U such that, for every $S \subset T$, P_S is defined.*

1. *If* $P(x, y) \neq 0, 1$ *for all* $x, y \in T$*, then for* $R \subset S \subset T$*:*

$$P_T(R) = P_S(R) P_T(S).$$

2. *If* $P(x, y) = 0$ *for some* $x, y \in T$*, then for every* $S \subset T$*:*

$$P_T(S) = P_{T-\{x\}}(S - \{x\}).$$

In what follows, we will drop the conditional dependency on class since it is not needed for our argument and leads to simpler notation. For brevity's sake, we shall also assume that case 2 never occurs (i.e. there are no elements that will always be picked over other elements with deterministic certainty). This relaxation has negligible influence and the proof for case 2 is pretty straightforward should it be of concern.

Lemma 1 *The (binary) multinomial event model agrees with the choice axiom.*

Proof Let us again consider three finite sets $R \subset S \subset T$ defined as follows:

$$T = \{X_1, X_2, \ldots, X_m\} \qquad S = \{X_1, X_2, \ldots X_k\} \qquad R = \{X_1, \ldots X_l\},$$

with $l < k$. We know from Eq. (11) that the probability of picking any choice within a set depends on the proportional weight with respect to the other elements from the set.

$$\begin{aligned} P_T(R) &= \sum_{j'=1}^{l} \frac{1 + \sum_{i=1}^{n} x_{i,j'} P(Y = y \mid \boldsymbol{x}_i)}{m + \sum_{i=1}^{n} \sum_{j=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)} \\ &= \frac{l + \sum_{i=1}^{n} \sum_{j=1}^{l} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{m + \sum_{i=1}^{n} \sum_{j=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)} \\ P_T(S) &= \frac{k + \sum_{i=1}^{n} \sum_{j=1}^{k} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{m + \sum_{i=1}^{n} \sum_{j=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}. \end{aligned}$$

The probability of picking a member of set R over a member in its complement set $\overline{R}$ in S is easily found as the sum of the individual multinomial probabilities of each element in R in the bag of choices S.

$$\begin{aligned} P_S(R) &= \sum_{j=1}^{l} \frac{1 + \sum_{i=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{k + \sum_{i=1}^{n} \sum_{j=1}^{k} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)} \\ &= \frac{l + \sum_{i=1}^{n} \sum_{j=1}^{l} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{k + \sum_{i=1}^{n} \sum_{j=1}^{k} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}. \end{aligned}$$

Luce's choice axiom follows trivially by multiplying these three equations.

Lemma 2 *The binary multi-variate Bernoulli event model violates the choice axiom.*

Proof We will provide a counter-example using the binary Bernoulli variant presented before. Let us consider three sets $R \subset S \subset T$ defined as follows:

$$T = \{X_1, X_2, \ldots, X_m\} \qquad S = \{X_1, \ldots, X_k\} \qquad R = \{X_1, \ldots, X_l\},$$

with $l < k$. We know from Eq. (9) that the probability of picking any choice depends on two counters, the number of non-zeroes for a particular choice (column) and the total number of instances (rows) in the data set, which without prior distribution leads to:

$$\begin{aligned} P_T(R) &= \sum_{j=1}^{l} P_T(X_j) \\ &= \sum_{j=1}^{l} \frac{1 + \sum_{i=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{2 + \sum_{i=1}^{n} P(Y = y \mid \boldsymbol{x}_i)} \\ P_T(S) &= \sum_{j=1}^{k} \frac{1 + \sum_{i=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{2 + \sum_{i=1}^{n} P(Y = y \mid \boldsymbol{x}_i)}. \end{aligned}$$

More importantly, the denominator in probabilities involving the subset S ($P_S(\cdot)$), potentially changes due to not all samples being in the member set under consideration. That is, there exist records for which we have no observed values in S, thus the total number of observed instances under consideration (n') is smaller than the original number of observations (n):

$$P_S(R) = \sum_{j=1}^{l} \frac{1 + \sum_{i=1}^{n'} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{2 + \sum_{i=1}^{n'} P(Y = y \mid \boldsymbol{x}_i)} \tag{15a}$$

$$= \sum_{j=1}^{l} \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left[\sum_{j=1}^{k} x_{i,j} > 0\right] x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{2 + \sum_{i=1}^{n} \mathbb{1}\left[\sum_{j=1}^{k} x_{i,j} > 0\right] P(Y = y \mid \boldsymbol{x}_i)} \tag{15b}$$

$$= \sum_{j=1}^{l} \frac{1 + \sum_{i=1}^{n} x_{i,j} P(Y = y \mid \boldsymbol{x}_i)}{2 + \sum_{i=1}^{n} \mathbb{1}\left[\sum_{j=1}^{k} x_{i,j} > 0\right] P(Y = y \mid \boldsymbol{x}_i)}. \tag{15c}$$

Note how we changed notation from n' to n in Eq. (15b) by taking under consideration only samples for which at least one of the $x_{i,j}$ is observed for at least one $X_j \in S$ by using the indicator operator which selects exactly these elements (indeed, the sum will be greater than zero if one of the elements is active). In the numerator we can drop this condition since any element for which $x_{i,j}$ is zero, would be cancelled out anyway due to the multiplication with $x_{i,j}$; this leads to the final formulation in Eq. (15c).

Clearly, for any non-trivial case containing non mutually exclusive variables:

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{k} x_{i,j} P(Y = y \mid \boldsymbol{x}_i) \right) > \sum_{i=1}^{n} \left(\mathbb{1} \left[\sum_{j=1}^{k} x_{i,j} > 0 \right] P(Y = y \mid \boldsymbol{x}_i) \right), \tag{16}$$

and thus after combination of the individual equations we arrive at:

$$P_T(R) < P_S(R) \cdot P_T(S).$$

Therefore, the Bernoulli event model does not satisfy the Choice Axiom.

Lemma 3 *The Wallenius event model is in agreement with the choice axiom.*

Proof Let us again consider three finite sets $R \subset S \subset T$ defined as follows:

$$T = \{X_1, X_2, \ldots, X_m\} \qquad S = T - \{X_1\} \qquad R = \{X_1\}$$

We know from Eq. (6) that the probability of picking any choice within a set, depends on the proportion of the weight of that choice with respect to the weight of the other elements from the set:

$$\begin{aligned} P_T(X_1) &= \frac{w_1}{\sum_{s \in S} w_s} \\ P_S(X_1) &= \frac{w_1}{\sum_{t \in T} w_t} \\ P_T(S) &= \frac{\sum_{s \in S} w_s}{\sum_{t \in T} w_t}. \end{aligned}$$

Just like for the multinomial event model, the first part of the axiom follows by filling in these values. Of course, matters get a little bit more complicated when multiple choices are selected since the permutations need to be taken into account. These will again factor out when multiplied with each other, leading to the confirmation of the axiom.

References

Asuncion, A., Newman, D. J. (2007). *UCI machine learning repository*. http://www.ics.uci.edu/~mlearn/MLRepository.html.

Bourdieu, P. (1984). *Distinction: A social critique of the judgement of taste*. Harvard University Press. http://books.google.com/books/about/Distinction.html?id=nVaS6gS9Jz4C&pgis=1.

Chesson, J. (1976). A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability*. http://www.jstor.org/stable/10.2307/3212535.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning Special Issue on Learning with Probabilistic Representations, 29*(2–3), 103 – 130. http://link.springer.com/article/10.1023/A:1007413511361.

Etter, J.-F., Le Houezec, J., & Perneger, T. (2003). 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. https://doi.org/10.1038/sj.npp.1300030.

Fantino, E., & Navarick, D. (1975). Recent developments in choice. In G. H. Bower (Ed.), *Psychology of learning & motivation* (p. 304). Academic Press. http://books.google.com/books?hl=en&lr=&id=o5LScJ9ecGUC&pgis=1.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. http://linkinghub.elsevier.com/retrieve/pii/S016786550500303X

Flach, P., & Lachiche, N. (2000). Decomposing probability distributions on structured individuals. In *Work-in-progress reports of the 10th international conference on inductive logic programming* (pp. 96–106). http://www.cs.bris.ac.uk/Publications/pub_master.jsp?id=1000485.

Flach, P., & Lachiche, N. (2004). Naive Bayesian classification of structured data. *Machine Learning*, *57*(3), 233–269. https://doi.org/10.1023/B:MACH.0000039778.69032.ab.

Fog, A. (2008). Calculation methods for Wallenius' noncentral hypergeometric distribution. *Communications in Statistics—Simulation and Computation*, *37*(2), 258–273. https://doi.org/10.1080/03610910701790269.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, *4*, 267–272. https://doi.org/10.1901/jeab.1961.4-267.

Junqué de Fortuny, E., Martens, D., & Provost, F. (2013). Predictive modeling with big data: Is bigger really better? *Big Data*, *1*(4), 215–226. https://doi.org/10.1089/big.2013.0037.

Kahaner, D., Moler, C., Nash, S. (1989). *Numerical methods and software*. Prentice Hall. ftp://ftp.math.utah.edu/pub/errata/kahaner.errata.

Kant, I. (1790). *The critique of judgement (Part one, the critique of aesthetic judgement)*. BiblioLife. http://www.amazon.com/The-Critique-Judgement-Part-Aesthetic/dp/1420926942.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). 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), 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Langley, P., Iba, W., & Thomposn, K. (1992). An analysis of Bayesian Classifers. In *Proceedings of the tenth national conference on artificial intelligence* (No. 415, pp. 223–228). https://pdfs.semanticscholar.org/1925/bacaa10b4ec83a0509132091bb79243b41b6.pdf.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Dover Publications. http://www.amazon.com/Individual-Choice-Behavior-Theoretical-Mathematics/dp/0486441369.

McCallum, A., Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In *AAAI workshop on learning for text categorization*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf.

Ng, A., Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. In *Advances in neural information processing systems*. https://papers.nips.cc/paper/2020-on-discriminative-vs-generative-classifiers-a-comparison-of-logistic-regression-and-naive-bayes.pdf.

Sapolsky, R., & Bonetta, L. (1997). *The trouble with testosterone: And other essays on the biology of the human predicament*. http://www.nature.com/nm/wilma/v3n8.870469132.html.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273. https://doi.org/10.1037/h0070288.

Wallenius, K. (1963). *Biased sampling: The non-central hypergeometric probability distribution*. Ph.D. thesis, Stanford University.

Ziegler, C., & McNee, S. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web* (pp. 22–32). http://dl.acm.org/citation.cfm?id=1060754.