
@inproceedings{Hardisty:Boyd-Graber:Resnik-2010,
Author = {Eric Hardisty and Jordan Boyd-Graber and Philip Resnik},
Url = {docs/adapted_naive_bayes.pdf},
Booktitle = {Empirical Methods in Natural Language Processing},
Location = {Cambridge, MA},
Year = {2010},
Title = {Modeling Perspective using Adaptor Grammars},
}

Downloaded from http://cs.colorado.edu/~jbg/docs/adapted_naive_bayes.pdf
Abstract

Strong indications of perspective can often come from collocations of arbitrary length; for example, someone writing get the government out of my X is typically expressing a conservative rather than progressive viewpoint. However, going beyond unigram or bigram features in perspective classification gives rise to problems of data sparsity. We address this problem using nonparametric Bayesian modeling, specifically adaptor grammars (Johnson et al., 2006). We demonstrate that an adaptive naïve Bayes model captures multiword lexical usages associated with perspective, and establishes a new state-of-the-art for perspective classification results using the Bitter Lemons corpus, a collection of essays about mid-east issues from Israeli and Palestinian points of view.

1 Introduction

Most work on the computational analysis of sentiment and perspective relies on lexical features. This makes sense, since an author’s choice of words is often used to express overt opinions (e.g. describing healthcare reform as idiotic or wonderful) or to frame a discussion in order to convey a perspective more implicitly (e.g. using the term death tax instead of estate tax). Moreover, it is easy and efficient to represent texts as collections of the words they contain, in order to apply a well known arsenal of supervised techniques (Laver et al., 2003; Mullen and Malouf, 2006; Yu et al., 2008).

At the same time, standard lexical features have their limitations for this kind of analysis. Such features are usually created by selecting some small n-gram size in advance. Indeed, it is not uncommon to see the feature space for sentiment analysis limited to unigrams. However, important indicators of perspective can also be longer (get the government out of my). Trying to capture these using standard machine learning approaches creates a problem, since allowing n-grams as features for larger n gives rise to problems of data sparsity.

In this paper, we employ nonparametric Bayesian models (Orbanz and Teh, 2010) in order to address this limitation. In contrast to parametric models, for which a fixed number of parameters are specified in advance, nonparametric models can “grow” to the size best suited to the observed data. In text analysis, models of this type have been employed primarily for unsupervised discovery of latent structure — for example, in topic modeling, when the true number of topics is not known (Teh et al., 2006); in grammatical inference, when the appropriate number of nonterminal symbols is not known (Liang et al., 2007); and in coreference resolution, when the number of entities in a given document is not specified in advance (Haghighi and Klein, 2007). Here we use them for supervised text classification.

Specifically, we use adaptor grammars (Johnson et al., 2006), a formalism for nonparametric Bayesian modeling that has recently proven useful in unsupervised modeling of phonemes (Johnson, 2008), grammar induction (Cohen et al., 2010), and named entity structure learning (Johnson, 2010), to make supervised naïve Bayes classification nonparametric in order to improve perspective modeling. Intuitively, naïve Bayes associates each class or label with a probability distribution over a fixed vocabulary. We introduce adaptive naïve Bayes (ANB), for which in principle the vocabulary can grow as needed to include collocations of arbitrary length, as determined...
by the properties of the dataset. We show that using adaptive naïve Bayes improves on state of the art classification using the Bitter Lemons corpus (Lin et al., 2006), a document collection that has been used by a variety of authors to evaluate perspective classification.

In Section 2, we review adaptor grammars, show how naïve Bayes can be expressed within the formalism, and describe how — and how easily — an adaptive naïve Bayes model can be created. Section 3 validates the approach via experimentation on the Bitter Lemons corpus. In Section 4, we summarize the contributions of the paper and discuss directions for future work.

2 Adapting Naïve Bayes to be Less Naïve

In this work we apply the adaptor grammar formalism introduced by Johnson, Griffiths, and Goldwater (Johnson et al., 2006). Adaptor grammars are a generalization of probabilistic context free grammars (PCFGs) that make it particularly easy to express non-parametric Bayesian models of language simply and readably using context free rules. Moreover, Johnson et al. provide an inference procedure based on Markov Chain Monte Carlo techniques that makes parameter estimation straightforward for all models that can be expressed using adaptor grammars.1 Variational inference for adaptor grammars has also been recently introduced (Cohen et al., 2010).

Briefly, adaptor grammars allow nonterminals to be rewritten to entire subtrees. In contrast, a non-terminal in a PCFG rewrites only to a collection of grammar symbols; their subsequent productions are independent of each other. For instance, a traditional PCFG might learn probabilities for the rewrite rule PP \( \rightarrow \) P NP. In contrast, an adaptor grammar can learn (or “cache”) the production PP \( \rightarrow \) (P up)(NP(DET a))(N tree)). It does this by positioning that the distribution over children for an adapted non-terminal comes from a Pitman-Yor distribution.

A Pitman-Yor distribution (Pitman and Yor, 1997) is a distribution over distributions. It has three parameters: the discount, \( a \), such that \( 0 \leq a < 1 \), the strength, \( b \), a real number such that \(-a < b\), and a probability distribution \( G_0 \) known as the base distribution. Adaptor grammars allow distributions over subtrees to come from a Pitman-Yor distribution with the PCFG’s original distribution over trees as the base distribution. The generative process for obtaining Draws from a distribution drawn from a Pitman-Yor distribution can be described by the “Chinese restaurant process” (CRP). We will use the CRP to describe how to obtain a distribution over observations composed of sequences of \( n \)-grams, the key to our model’s ability to capture perspective-bearing \( n \)-grams.

Suppose that we have a base distribution \( \Omega \) that is some distribution over all sequences of words (the exact structure of such a distribution is unimportant; such a distribution will be defined later in Table 1). Suppose further we have a distribution \( \phi \) drawn from \( PY(a, b, \Omega) \), and we wish to draw a series of observations \( w \) from \( \phi \). The CRP gives us a generative process for doing those draws from \( \phi \), marginalizing out \( \phi \). Following the restaurant metaphor, we imagine the \( i \)th customer in the series entering the restaurant to take a seat at a table. The customer sits by making a choice that determines the value of the \( n \)-gram \( w_i \) for that customer: she can either sit at an existing table or start a new table of her own.2

If she sits at a new table \( j \), that table is assigned a draw \( y_j \) from the base distribution, \( \Omega \); note that, since \( \Omega \) is a distribution over \( n \)-grams, \( y_j \) is an \( n \)-gram. The value of \( w_i \) is therefore assigned to be \( y_j \), and \( y_j \) becomes the sequence of words assigned to that new table. On the other hand, if she sits at an existing table, then \( w_i \) simply takes the sequence of words already associated with that table (assigned as above when it was first occupied).

The probability of joining an existing table \( j \), with \( c_j \) patrons already seated at table \( j \), is \( \frac{c_j - a}{c_j + b} \), where \( c \) is the number of patrons seated at all tables:

\[
c_j = \sum_{j'} c_{j'}
\]

The probability of starting a new table is \( \frac{b + t a}{c + t} \), where \( t \) is the number of tables presently occupied.

Notice that \( \phi \) is a distribution over the same space as \( \Omega \), but it can drastically shift the mass of the distribution, compared with \( \Omega \), as more and more pa-

---

1And, better still, they provide code that implements the inference algorithm; see http://www.cog.brown.edu/ mj/Software.htm.

2Note that we are abusing notation by allowing \( w_i \) to correspond to a word sequence of length \( \geq 1 \) rather than a single word.
trons are seated at tables. However, there is always a chance of drawing from the base distribution, and therefore every word sequence can also always be drawn from $\phi$.

In the next section we will write a naïve Bayes-like generative process using PCFGs. We will then use the PCFG distribution as the base distribution for a Pitman-Yor distribution, adapting the naïve Bayes process to give us a distribution over $n$-grams, thus learning new language substructures that are useful for modeling the differences in perspective.

### 2.1 Classification Models as PCFGs

Naïve Bayes is a venerable and popular mechanism for text classification (Lewis, 1998). It posits that there are $K$ distinct categories of text — each with a distinct distribution over words — and that every document, represented as an exchangeable bag of words, is drawn from one (and only one) of these distributions. Learning the per-category word distributions and global prevalence of the classes is a problem of posterior inference which can be approached using a variety of inference techniques (Lowd and Domingos, 2005).

More formally, naïve Bayes models can be expressed via the following generative process:\footnote{Here $\alpha$ and $\lambda$ are hyperparameters used to specify priors for the class distribution and classes’ word distributions, respectively; $\alpha$ is a symmetric $K$-dimensional vector where each element is $\pi$. $N_d$ is the length of document $d$. Resnik and Hardisty (2010) provide a tutorial introduction to the naïve Bayes generative process and underlying concepts.}

1. Draw a global distribution over classes $\theta \sim \text{Dir} (\alpha)$
2. For each class $i \in \{1, \ldots, K\}$, draw a word distribution $\phi_i \sim \text{Dir} (\lambda)$
3. For each document $d \in \{1, \ldots, M\}$:
   (a) Draw a class assignment $z_d \sim \text{Mult} (\theta)$
   (b) For each word position $n \in \{1, \ldots, N_d\}$, draw $w_{d,n} \sim \text{Mult} (\phi_{z_d})$

A variant of the naïve Bayes generative process can be expressed using the adaptor grammar formalism (Table 1). The left hand side of each rule represents a nonterminal which can be expanded, and the right hand side represents the rewrite rule. The rightmost indices show replication; for instance, there are $|V|$ rules that allow $\text{WORD}_i$ to rewrite to each word in the vocabulary. One can assume a symmetric Dirichlet prior of $\text{Dir} (1)$ over the production choices unless otherwise specified — as with the $\text{DOC}_d$ production rule above, where a sparse prior is used.

Notice that the distribution over expansions for $\text{WORD}_i$ corresponds directly to $\phi_i$ in Figure 1(a). There are, however, some differences between the model that we have described above and the standard naïve Bayes model depicted in Figure 1(a). In particular, there is no longer a single choice of class per document; each sentence is assigned a class. If the distribution over per-sentence labels is sparse (as it is above for $\text{DOC}_d$), this will closely approximate naïve Bayes, since it will be very unlikely for the sentences in a document to have different labels. A non-sparse prior leads to behavior more like models that allow parts of a document to express sentiment or perspective differently.

### 2.2 Moving Beyond the Bag of Words

The naïve Bayes generative distribution posits that when writing a document, the author selects a distribution of categories $z_d$ for the document from $\theta$. The author then generates words one at a time: each word is selected independently from a flat multinomial distribution $\phi_{z_d}$ over the vocabulary.

However, this is a very limited picture of how text is related to underlying perspectives. Clearly words are often connected with each other as collocations, and, just as clearly, extending a flat vocabulary to include bigram collocations does not suffice, since sometimes relevant perspective-bearing phrases are longer than two words. Consider phrases like health care for all or government takeover of health care, connected with progressive and conservative positions, respectively, during the national debate on healthcare reform. Simply applying naïve Bayes, or any other model, to a bag of $n$-grams for high $n$ is
going to lead to unworkable levels of data sparsity; a model should be flexible enough to support both unigrams and longer phrases as needed.

Following Johnson (2010), however, we can use adaptor grammars to extend naïve Bayes flexibly to include richer structure like collocations when they improve the model, and not including them when they do not. This can be accomplished by introducing adapted nonterminal rules: in a revised generative process, the author can draw from Pitman-Yor distribution whose base distribution is over word sequences of arbitrary length. Thus in a setting where, say, $K = 2$, and our two classes are PROGRESSIVE and CONSERVATIVE, the sequence health care for all might be generated as a single unit for the progressive perspective, but in the conservative perspective the same sequence might be generated as three separate draws: health care, for, all. Such a model is presented in Figure 1(b). Note the following differences between Figures 1(a) and 1(b):

- $z_d$ selects which Pitman-Yor distribution to draw from for document $d$.
- $\phi_i$ is the distribution over $n$-grams that comes from the Pitman-Yor distribution.
- $W_{d,n}$ represents an $n$-gram draw from $\phi_i$.
- $a, b$ are the Pitman-Yor strength and discount parameters.
- $\Omega$ is the Pitman-Yor base distribution with $\tau$ as its uniform hyperparameter.

Returning to the CRP metaphor discussed when we introduced the Pitman-Yor distribution, there are two restaurants, one for the PROGRESSIVE distribution and one for the CONSERVATIVE distribution. Health care for all has its own table in the PROGRESSIVE restaurant, and enough people are sitting at it to make it popular. There is no such table in the CONSERVATIVE restaurant, so in order to generate those words, the phrase health care for all would need to come from a new table; however, it is more easily explained by three customers sitting at three existing, popular tables: health care, for, and all.

We follow the convention of Johnson (2010) by writing adapted nonterminals as underlined. The grammar for adaptive naïve Bayes is shown in Table 2. The adapted COLLOC$_i$ rule means that every time we need to generate that nonterminal, we are actually drawing from a distribution drawn from a Pitman-Yor distribution. The distribution over the possible yields of the WORDS$_i$ rule serves as the base distribution.

Given this generative process for documents, we can now use statistical inference to uncover the posterior distribution over the latent variables, thus discovering the tables and seating assignments of our metaphorical restaurants that each cater to a specific perspective filled with tables populated by words and $n$-grams.

The model presented in Table 2 is the most straightforward way of extending naïve Bayes to collocations. For completeness, we also consider the alternative of using a shared base distribution rather than distinguishing the base distributions of the two classes.
Two of the authors are guests, one from each perspective, for a total of four essays on each topic per week. We chose this corpus to allow us to directly compare our results with Greene and Resnik’s (2009) Observable Proxies for Underlying Semantics (OPUS) features and Lin et al.’s Latent Sentence Perspective Model (LSPM). The classification goal for this corpus is to label each document with the perspective of its author, either Israeli or Palestinian.

Consistent with prior work, we prepared the corpus by dividing it into two groups, one group containing all of the essays written by the regular site contributors, which we call the Editor set, and one group comprised of all the essays written by the guest contributors, which we call the Guest set. Similar to the above mentioned prior work, we perform classification using one group as training data and the other as test data and perform two folds of classification. The overall experimental setup and corpus preparation process is presented in Figure 3.
### 3.2 Experimental Setup

The vocabulary generator determines the vocabulary used by a given experiment by converting the training set to lower case, stemming with the Porter stemmer, and filtering punctuation. We remove from the vocabulary any words that appeared in only one document regardless of frequency within that document, words with frequencies lower than a threshold, and stop words.\(^6\) The vocabulary is then passed to a grammar generator and a corpus filter.

The grammar generator uses the vocabulary to generate the terminating rules of the grammar from the ANB grammar presented in Tables 2 and 3. The corpus filter takes in a set of documents and replaces all words not in the vocabulary with “out of vocabulary” markers. This process ensures that in all experiments the vocabulary is composed entirely of words from the training set. After the groups have been filtered, the group used as the test set has its labels removed. The test and training set are then sent, along with the grammar, into the adaptor grammar inference engine.

Each experiment ran for 3000 iterations. For the runs where adaptation was used we set the initial Pitman-Yor\(\alpha\) and\(\beta\) parameters to 0.01 and 10 respectively, then slice sample (Johnson and Goldwater, 2009).

We use the resulting sentence parses for classification. By design of the grammar, each sentence’s words will belong to one and only one distribution. We identify that distribution from each of the test set sentence parses and use it as the sentence level classification for that particular sentence. We then use majority rule on the individual sentence classifications in a document to obtain the document classification. (In most cases the sentence-level assignments are overwhelmingly dominated by one class.)

### 3.3 Results and Analysis

Table 4 gives the results and compares to prior work. The support vector machine (SVM), NB-B and LSPM results are taken directly from Lin\ et\ al. (2006). NB-B indicates naïve Bayes with full Bayesian inference. LSPM is the Latent Sentence Perspective Model, also from Lin\ et\ al. (2006). OPUS results are taken from Greene and Resnik (2009). Briefly, OPUS features are generated from observable grammatical relations that come from dependency parses of the corpus. Use of these features provided the best classification accuracy for this task prior to this work. ANB* refers to the grammar from Table 2, but with adaptation disabled. The reported accuracy values for ANB*, ANB with a common base distribution (see Table 3), and ANB with separate base distributions (see Table 2) are the mean values from five separate sampling chains. Bold face indicates statistical significance (\(p < 0.05\)) by unpaired t-test between the reported value and ANB*.

Table 5 presents some data on adapted rules and Resnik (2009). Briefly, OPUS features are generated from observable grammatical relations that come from dependency parses of the corpus. Use of these features provided the best classification accuracy for this task prior to this work. ANB* refers to the grammar from Table 2, but with adaptation disabled. The reported accuracy values for ANB*, ANB with a common base distribution (see Table 3), and ANB with separate base distributions (see Table 2) are the mean values from five separate sampling chains. Bold face indicates statistical significance (\(p < 0.05\)) by unpaired t-test between the reported value and ANB*.

Consistent with all prior work on this corpus we found that the classification accuracy for training on editors and testing on guests was lower than the other direction since the larger number of editors in the guest set allows for greater generalization. The difference between ANB* and ANB with a common base distribution is not statistically significant. Also of note is that the classification accuracy improves for testing on Guests when the ANB grammar is allowed to adapt and a separate base distribution is used for the two classes (88.28% versus 84.98% without adaptation).

Table 5 presents some data on adapted rules

<table>
<thead>
<tr>
<th>Training Set</th>
<th>Test Set</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guests</td>
<td>Editors</td>
<td>SVM</td>
<td>88.22</td>
</tr>
<tr>
<td>Guests</td>
<td>Editors</td>
<td>NB-B</td>
<td>93.46</td>
</tr>
<tr>
<td>Guests</td>
<td>Editors</td>
<td>LSPM</td>
<td>94.93</td>
</tr>
<tr>
<td>Guests</td>
<td>Editors</td>
<td>OPUS</td>
<td>97.64</td>
</tr>
<tr>
<td>Guests</td>
<td>Editors</td>
<td>ANB*</td>
<td>99.32</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>ANB Sep</td>
<td>99.87</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>SVM</td>
<td>81.48</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>NB-B</td>
<td>85.85</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>LSPM</td>
<td>86.99</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>OPUS</td>
<td>85.86</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>ANB*</td>
<td>84.98</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>ANB Com</td>
<td>82.76</td>
</tr>
<tr>
<td>Editors</td>
<td>Guests</td>
<td>ANB Sep</td>
<td><strong>88.28</strong></td>
</tr>
</tbody>
</table>

Table 4: Classification results. ANB* indicates the same grammar as Adapted Naïve Bayes, but with adaptation disabled. Com and Sep refer to whether the base distribution was common to both classes or separate.

\(^6\)In these experiments, a frequency threshold of 4 was selected prior to testing.
Learned once inference is complete. The column labeled "unique unigrams cached" indicates the number of unique unigrams that appear on the right hand side of the adapted rules. Similarly, "unique n-grams cached" indicates the number of unique n-grams that appear on the right hand side of the adapted rules. The rightmost column indicates the percentage of terms from the group vocabulary that appear on the right hand side of adapted rules as unigrams. Values less than 100% indicate that the remaining vocabulary terms are cached in n-grams. As the table shows, a significant number of the rules learned during inference are n-grams of various sizes.

Inspection of the captured bigrams showed that it captured sequences that a human might associate with one perspective over the other. Table 6 lists just a few of the more charged bigrams that were captured in the adapted rules.

Table 6: Charged bigrams captured by the framework.

<table>
<thead>
<tr>
<th>Israeli</th>
<th>Palestinian</th>
</tr>
</thead>
<tbody>
<tr>
<td>zionist dream</td>
<td>american jew</td>
</tr>
<tr>
<td>zionist state</td>
<td>achieve freedom</td>
</tr>
<tr>
<td>zionist movement</td>
<td>palestinian freedom</td>
</tr>
<tr>
<td>american leadership</td>
<td>support palestinian</td>
</tr>
<tr>
<td>american victory</td>
<td>palestinian suffer</td>
</tr>
<tr>
<td>abandon violence</td>
<td>palestinian territory</td>
</tr>
<tr>
<td>freedom (of the) press</td>
<td>palestinian statehood</td>
</tr>
<tr>
<td>palestinian violence</td>
<td>palestinian refugee</td>
</tr>
</tbody>
</table>

The same number of times but appears only 18 times in the corpus. In other words, the sequence people go is more easily explained by two sequential unigrams, not a bigram. The ratio of cache use counts to raw bigrams gives a measure of strength of collocation between the terms of the n-gram. We conjecture that the rareness of caching for \( n > 2 \) is a function of the small corpus size. Also of note is the improvement in performance of ANB* over NB-B when training on guests, which we suspect is due to our use of sampled versus fixed hyperparameters.

4 Conclusions

In this paper, we have applied adaptor grammars in a supervised setting to model lexical properties of text and improve document classification according to perspective, by allowing nonparametric discovery of collocations that aid in perspective classification. The adaptive naïve Bayes model improves on state of the art supervised classification performance in head-to-head comparisons with previous approaches.

Although there have been many investigations on the efficacy of using multiword collocations in text classification (Bekkerman and Allan, 2004), usually such approaches depend on a preprocessing step such as computing tf-idf or other measures of frequency based on either word bigrams (Tan et al., 2002) or character n-grams (Raskutti et al., 2001). In contrast, our approach combines phrase discovery with the probabilistic model of the text. This allows for the collocation selection and classification to be expressed in a single model, which can then be extended later; it also is truly generative, as compared with feature induction and selection algorithms that either under- or over-generate data.

There are a number of interesting directions in which to take this research. As Johnson et al. (2006) argue, and as we have confirmed here, the adaptor
The second column indicates the raw count of that bigram in the Guests or Editors group.

Table 7: Most frequently used cached bigrams. The first column in each section is the number of times that bigram was used as a cached rule. The second column indicates the raw count of that bigram in the Guests or Editors group.

Running this grammar did not produce improvements consistent with those reported by Lin et al. We plan to investigate this further, and a natural follow-on would be to experiment with adaptation for this variety of latent structure, to produce an adapted LSPM-like model analogous to adaptive naïve Bayes.

Acknowledgments

This research was funded in part by the Army Research Laboratory through ARL Cooperative Agreement W911NF-09-2-0072 and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies.
of ARL, IARPA, the ODNI, or the U.S. Government. The authors thank Mark Johnson and the anonymous reviewers for their helpful comments and discussions. We are particularly grateful to Mark Johnson for making his adaptor grammar code available.

References


