Using Vocabulary Knowledge in Bayesian Multinomial Estimation

Thomas L. Griffiths & Joshua B. Tenenbaum
Department of Psychology
Stanford University, Stanford, CA 94305
{gruffydd,jbt}@psych.stanford.edu

Abstract

Estimating the parameters of sparse multinomial distributions is an important component of many statistical learning tasks. Recent approaches have used uncertainty over the vocabulary of symbols in a multinomial distribution as a means of accounting for sparsity. We present a Bayesian approach that allows weak prior knowledge, in the form of a small set of approximate candidate vocabularies, to be used to dramatically improve the resulting estimates. We demonstrate these improvements in applications to text compression and estimating distributions over words in newsgroup data.

1 Introduction

Sparse multinomial distributions arise in many statistical domains, including natural language processing and graphical models. Consequently, a number of approaches to parameter estimation for sparse multinomial distributions have been suggested [3]. These approaches tend to be domain-independent: they make little use of prior knowledge about a specific domain. In many domains where multinomial distributions are estimated there is often at least weak prior knowledge about the potential structure of distributions, such as a set of hypotheses about restricted vocabularies from which the symbols might be generated. Such knowledge can be solicited from experts or obtained from unlabeled data. We present a method for Bayesian parameter estimation in sparse discrete domains that exploits this weak form of prior knowledge to improve estimates over knowledge-free approaches.

1.1 Bayesian parameter estimation for multinomial distributions

Following the presentation in [4], we consider a language $\Sigma$ containing $L$ distinct symbols. A multinomial distribution is specified by a parameter vector $\theta = \langle \theta_1, \ldots, \theta_L \rangle$, where $\theta_i$ is the probability of an observation being symbol $i$. Consequently, we have the constraints that $\sum_{i=1}^{L} \theta_i = 1$ and $\theta_i \geq 0, i = 1, \ldots, L$.

The task of multinomial estimation is to take a data set $D$ and produce a vector $\theta$ that results in a good approximation to the distribution that produced $D$. In this case, $D$ consists of $N$ independent observations $x^1, \ldots, x^N$ drawn from the distribution to be estimated, which can be summarized by the statistics $N_i$ specifying the number of times the $i$th symbol occurs in the data. $D$ also determines the set $\Sigma^0$
of symbols that occur in the data.

Stated in this way, multinomial estimation involves predicting the next observation based on the data. Specifically, we wish to calculate \( P(x^{N+1}|D) \). The Bayesian estimate for this probability is given by

\[
P_L(x^{N+1}|D) = \int P(x^{N+1}|\theta)P(\theta|D)d\theta
\]

where \( P(x^{N+1}|\theta) \) follows from the multinomial distribution corresponding to \( \theta \). The posterior probability \( P(\theta|D) \) can be obtained via Bayes rule

\[
P(\theta|D) \propto P(D|\theta)P(\theta) = P(\theta)\prod_{i=1}^{L} \theta_{N_i}
\]

where \( P(\theta) \) is the prior probability of a given \( \theta \).

Laplace used this method with a uniform prior over \( \theta \) to give the famous “law of succession” [6]. A more general approach is to assume a Dirichlet prior over \( \theta \), which is conjugate to the multinomial distribution and gives

\[
P(X^{N+1} = i|D) = \frac{N_i + \alpha_i}{N + \sum_{j=1}^{L} \alpha_j}
\]

where the \( \alpha_i \) are the hyperparameters of the Dirichlet distribution. Different estimates are obtained for different choices of the \( \alpha_i \), with most approaches making the simplifying assumption that \( \alpha_i = \alpha \) for all \( i \). Laplace’s law results from \( \alpha = 1 \). The case with \( \alpha = 0.5 \) is the Jeffreys-Perks law or Expected Likelihood Estimation [2] [5] [9], while using arbitrary \( \alpha \) is Lidstone’s law [7].

### 1.2 Estimating sparse multinomial distributions

Several authors have extended the Bayesian approach to sparse multinomial distributions, in which only a restricted vocabulary of symbols are used, by maintaining uncertainty over these vocabularies. In [10], Ristad uses assumptions about the probability of strings based upon different vocabularies to give the estimate

\[
P_R(X^{N+1} = i|D) = \begin{cases} 
\frac{(N_i + 1)/(N + L)}{L} & \text{if } k^0 = L \\
\frac{(N_i + 1)(N + 1 - k^0)/(N^2 + N + 2k^0)}{k^0(k^0 + 1)/(L - k^0)(N^2 + N + 2k^0)} & \text{if } k^0 < L \land N_i > 0 \\
k^0 & \text{otherwise}
\end{cases}
\]

where \( k^0 = |\Sigma^0| \) is the size of the smallest vocabulary consistent with the data.

A different approach is taken by Friedman and Singer in [4], who point out that Ristad’s method is a special case of their framework. Friedman and Singer consider the vocabulary \( V \subseteq \Sigma \) to be a random variable, allowing them to write

\[
P(X^{N+1} = i|D) = \sum_{V} P(X^{N+1} = i|V, D)P(V|D)
\]

where \( P(X^{N+1} = i|V, D) \) results from a Dirichlet prior over the symbols in \( V \),

\[
P(X^{N+1} = i|V, D) = \begin{cases} 
\frac{N_i + \alpha}{N + |V|\alpha} & \text{if } i \in V \\
0 & \text{otherwise}
\end{cases}
\]

and by Bayes’ rule and the properties of Dirichlet priors

\[
P(V|D) \propto P(D|V)P(V) \]

\[
= \begin{cases} 
\frac{\Gamma(|V|\alpha)}{\Gamma(|V|\alpha)}\prod_{i \in \Sigma^0} \frac{\Gamma(N_i + \alpha)}{\Gamma(\alpha)}P(V) & \Sigma^0 \subseteq V \\
0 & \text{otherwise}
\end{cases}
\]
Friedman and Singer assume a hierarchical prior over $V$, such that all vocabularies of cardinality $k$ are given equal probability, namely $P(S = k) = (1/k)^{N}$, where $P(S = k)$ is the probability that the size of the vocabulary $|V|$ is $k$. It follows that if $i \in \Sigma^0$, $P(X^{N+1} = i|D) = \sum_k \frac{N+k}{N+k-\alpha} P(S = k|D)$. If $i \notin \Sigma^0$, it is necessary to estimate the proportion of $V$ that contain $i$ for a given $k$. The simplified result is

$$P_F(X^{N+1} = i|D) = \left\{ \begin{array}{ll}
\frac{N+k}{N+k-\alpha} C(D, L) & \text{if } i \in \Sigma^0 \\
\frac{N+k}{N+k-\alpha} (1 - C(D, L)) & \text{otherwise}
\end{array} \right. \quad (5)$$

where

$$C(D, L) = \sum_{k=0}^{L} \frac{N+k}{N+k-\alpha} m_k$$

with $m_k = P(S = k) \frac{k!}{(k-\alpha)!} \frac{\Gamma(k+\alpha)}{\Gamma(N+k+\alpha)}$.

2 Making use of weak prior knowledge

Friedman and Singer assume a prior that gives equal probability to all vocabularies of a given cardinality. However, many real-world tasks provide limited knowledge about the structure of distributions that we can build into our methods for parameter estimation. In the context of sparse multinomial estimation, one instance of such knowledge the importance of specific vocabularies. For example, in predicting the next character in a file, our predictions could be facilitated by considering the fact that most files either use a vocabulary consisting of ASCII printing characters (such as text files), or all possible characters (such as object files). This kind of structural knowledge about a domain is typically easier to solicit from experts than accurate distributional information, and forms a valuable informational resource.

If we have this kind of prior knowledge, we can restrict our attention to a subset of the $2^L$ possible vocabularies. In particular, we can specify a set of vocabularies $\mathcal{V}$ which we consider as hypotheses for the vocabulary used in producing $D$, where the elements of $\mathcal{V}$ are specified by our knowledge of the domain. This stands as a compromise between Friedman and Singer's approach, in which $\mathcal{V}$ consists of all vocabularies, and traditional Bayesian parameter estimation as represented by Equation 1, in which $\mathcal{V}$ consists of only the vocabulary containing all words. To do this, we explicitly evaluate the sum given in Equation 2, where the sum over $\mathcal{V}$ includes all $\mathcal{V} \in \mathcal{V}$. This sum remains tractable when $\mathcal{V}$ is a small subset of the possible vocabularies, and the efficiency is aided by the fact that $P(D|\mathcal{V})$ shares common terms across all $\mathcal{V}$ which can cancel in normalization.

The intuition behind this approach is that it attempts to classify the target distribution as using one of a known set of vocabularies, where the vocabularies are obtained either from experts or from unlabeled data. Applying standard Bayesian multinomial estimation within this vocabulary gives enough flexibility for the method to capture a range of distributions, while making use of our weak prior knowledge.

2.1 An illustration: Text compression

Text compression is an effective test of methods for multinomial estimation. Adaptive coding can be performed by specifying a method for calculating a distribution over the probability of the next byte in a file based upon the preceding bytes [1]. The extent to which the file is compressed depends upon the quality of these predictions. To illustrate the utility of including prior knowledge, we follow Ristad in using the Calgary text compression corpus [1]. This corpus consists of 19 files of
Table 1: Text compression lengths (in bytes) on the Calgary corpus

<table>
<thead>
<tr>
<th>file</th>
<th>size</th>
<th>( k^0 )</th>
<th>( NH(N_i/N) )</th>
<th>( P_V )</th>
<th>( P_F )</th>
<th>( P_R )</th>
<th>( P_L )</th>
<th>( P_I )</th>
</tr>
</thead>
<tbody>
<tr>
<td>bib</td>
<td>111261</td>
<td>81</td>
<td>72330</td>
<td>78</td>
<td>89</td>
<td>92</td>
<td>269</td>
<td>174</td>
</tr>
<tr>
<td>book1</td>
<td>768771</td>
<td>82</td>
<td>435043</td>
<td>219</td>
<td>105</td>
<td>116</td>
<td>352</td>
<td>219</td>
</tr>
<tr>
<td>book2</td>
<td>610856</td>
<td>96</td>
<td>365952</td>
<td>94</td>
<td>115</td>
<td>124</td>
<td>329</td>
<td>212</td>
</tr>
<tr>
<td>geo</td>
<td>102400</td>
<td>256</td>
<td>72274</td>
<td>161</td>
<td>162</td>
<td>165</td>
<td>165</td>
<td>161</td>
</tr>
<tr>
<td>news</td>
<td>377109</td>
<td>98</td>
<td>244632</td>
<td>89</td>
<td>113</td>
<td>116</td>
<td>304</td>
<td>201</td>
</tr>
<tr>
<td>obj1</td>
<td>21504</td>
<td>256</td>
<td>159899</td>
<td>126</td>
<td>127</td>
<td>129</td>
<td>129</td>
<td>126</td>
</tr>
<tr>
<td>obj2</td>
<td>246814</td>
<td>256</td>
<td>193144</td>
<td>182</td>
<td>184</td>
<td>189</td>
<td>189</td>
<td>182</td>
</tr>
<tr>
<td>paper1</td>
<td>53161</td>
<td>95</td>
<td>31113</td>
<td>71</td>
<td>94</td>
<td>100</td>
<td>236</td>
<td>156</td>
</tr>
<tr>
<td>paper2</td>
<td>62379</td>
<td>91</td>
<td>47280</td>
<td>75</td>
<td>94</td>
<td>105</td>
<td>259</td>
<td>167</td>
</tr>
<tr>
<td>paper3</td>
<td>46326</td>
<td>84</td>
<td>27132</td>
<td>70</td>
<td>85</td>
<td>92</td>
<td>238</td>
<td>154</td>
</tr>
<tr>
<td>paper4</td>
<td>13286</td>
<td>80</td>
<td>7806</td>
<td>58</td>
<td>72</td>
<td>79</td>
<td>190</td>
<td>126</td>
</tr>
<tr>
<td>paper5</td>
<td>11954</td>
<td>91</td>
<td>7376</td>
<td>57</td>
<td>79</td>
<td>83</td>
<td>181</td>
<td>122</td>
</tr>
<tr>
<td>paper6</td>
<td>38105</td>
<td>93</td>
<td>23861</td>
<td>68</td>
<td>90</td>
<td>95</td>
<td>223</td>
<td>149</td>
</tr>
<tr>
<td>pic</td>
<td>513216</td>
<td>159</td>
<td>77636</td>
<td>205</td>
<td>162</td>
<td>216</td>
<td>323</td>
<td>205</td>
</tr>
<tr>
<td>proge</td>
<td>39611</td>
<td>92</td>
<td>25743</td>
<td>68</td>
<td>89</td>
<td>91</td>
<td>222</td>
<td>150</td>
</tr>
<tr>
<td>prog2</td>
<td>71646</td>
<td>87</td>
<td>42720</td>
<td>74</td>
<td>91</td>
<td>97</td>
<td>253</td>
<td>164</td>
</tr>
<tr>
<td>prosp</td>
<td>45379</td>
<td>89</td>
<td>30052</td>
<td>71</td>
<td>89</td>
<td>94</td>
<td>236</td>
<td>155</td>
</tr>
<tr>
<td>trans</td>
<td>93695</td>
<td>99</td>
<td>64800</td>
<td>169</td>
<td>101</td>
<td>105</td>
<td>252</td>
<td>169</td>
</tr>
</tbody>
</table>

The files include BibTeX source (bib), formatted English text (book*, paper*), geological data (geo), newsgroup articles (news), object files (obj*), a bit-mapped picture (pic), programs in three different languages (prog*) and a terminal transcript (trans). The task was to estimate the distribution from which characters in the file were drawn based upon the first \( N \) characters and thus predict the \( N + 1 \)st character. Performance was measured in terms of the length of the resulting file, where the contribution of the \( N + 1 \)st character to the length is \( \log_2 P(\alpha^{N+1} / D) \). The results are expressed as the number of bits required to encode the file relative to the empirical entropy \( NH(N_i/N) \) as assessed by Ristad [10].

Results are shown in Table 1. \( P_V \) is the restricted vocabulary model outlined above, with \( V \) consisting of just two hypotheses: one corresponding to binary files, containing all 256 characters, and one consisting of a 107 character vocabulary representing formatted English. The latter vocabulary was estimated from 5MB of English text, C code, BibTeX source, and newsgroup data from outside the Calgary corpus. \( P_F \) is Friedman and Singer's method. For both of these approaches, \( \alpha \) was set to 0.5, to allow direct comparison to the Jeffreys-Perks law, \( P_J \). \( P_R \) and \( P_L \) are Ristad's and Laplace's laws respectively. \( P_I \) outperformed the other methods on all files based upon English text, bar book1, and all files using all 256 symbols\(^1\). The high performance followed from rapid classification of these files as using the appropriate vocabulary in \( V \). When the vocabulary included all symbols \( P_V \) performed as \( P_I \), which gave the best predictions for these files.

\(^{1}\)A number of excellent techniques for text compression exist that outperform all of those presented here. We have not included these techniques for comparison because our interest is in using text compression as a means of assessing estimation procedures, rather than as an end in itself. We thus consider only methods for multinomial estimation as our comparison group.
2.2 Maintaining uncertainty in vocabularies

The results for book1 illustrate a weakness of the approach outlined above. The file length for \( P_V \) is higher than those for \( P_F \) and \( P_H \), despite the fact that the file uses a text-based vocabulary. This file contains two characters that were not encountered in the data used to construct \( V \). These characters caused \( P_V \) to default to the unrestricted vocabulary of all 256 characters. From that point \( P_V \) corresponded to \( P_J \), which gave poor results on this file.

This behavior results from the assumption that the candidate vocabularies in \( V \) are completely accurate. Since in many cases the knowledge that informs the vocabularies in \( V \) may be imperfect, it is desirable to allow for uncertainty in vocabularies. This uncertainty will be reflected in the fact that symbols outside \( V \) are expected to occur with a vocabulary-specific probability \( \epsilon_v \),

\[
P(X^{N+1} = i | V, D) = \begin{cases} 
(1 - (L - |V|/\epsilon_V))^{N_V + \alpha} / \epsilon_V & \text{if } i \in V \\
\epsilon_V & \text{otherwise}
\end{cases}
\]

where \( N_V = \sum_{i \in V} N_i \). It follows that

\[
P(D | V) = (1 - (L - |V|/\epsilon_V))^{N_V} / \epsilon_V \prod_{i \in V} \Gamma(N_i + \alpha) / \Gamma(N_i)
\]

which replaces Equations 3-4 in specifying \( P_V \).

When \( V \) is determined by the judgments of domain experts, \( \epsilon_V \) is the probability that an unmentioned word actually belongs to a particular vocabulary. While it may not be the most efficient use of such data, the \( V \in V \) can also be estimated from some form of unlabeled data. In this case, Friedman and Singer’s approach provides a means of setting \( \epsilon_V \). Friedman and Singer explicitly calculate the probability that an unseen word is in \( V \) based upon a dataset: from the second condition of Equation 5, we find that we should set \( \epsilon_V = \frac{1}{1 + \prod_{i \in V} (1 - C(D, L))} \). We use this method below.

3 Bayesian parameter estimation in natural language

Statistical natural language processing often uses sparse multinomial distributions over large vocabularies of words. In different contexts, different vocabularies will be used. By specifying a basis set of vocabularies, we can perform parameter estimation by classifying distributions according to their vocabulary. This idea was examined using data from 20 different Usenet newsgroups. This dataset is commonly used in testing text classification algorithms (e.g. [8]). Ten newsgroups were used to estimate a set of vocabularies \( V \) with corresponding \( \epsilon_V \). These vocabularies were used in estimating multinomial distributions on these newsgroups and ten others.

The dataset was 20news-18827, which consists of the 20newsgroups data with headers and duplicates removed, and was preprocessed to remove all punctuation, capitalization, and distinct numbers. The articles in each of the 20 newsgroups were then divided into three sets. The first 500 articles from ten newsgroups were used to estimate the candidate vocabularies \( V \) and uncertainty parameters \( \epsilon_V \). Articles 501-700 for all 20 newsgroups were used as training data for multinomial estimation. Articles 701-900 for all 20 newsgroups were used as testing data. Following [8], a dictionary was built up by running over the 13,000 articles resulting from this division, and all words that occurred only once were mapped to an “unknown” word. The resulting dictionary contained \( L = 54309 \) words.

As before, the restricted vocabulary method \( (P_V) \), Friedman and Singer’s method \( (P_F) \), and Ristad’s \( (P_R) \), Laplace’s \( (P_L) \) and the Jeffreys-Perks \( (P_J) \) laws were ap-